N^o d'ordre:

THÈSE

présentée

devant l'université de Rennes 1

pour obtenir

le grade de : Docteur de l'université de Rennes 1
Mention Traîtement du signal

par

François Tonnin

Équipes d'accueil : TEXMEX-TEMICS
École doctorale : MATISSE
Composante universitaire : IRISA-IFSIC

Titre de la thèse :

Description locale d'images fixes dans le domaine compressé

Soutenue le devant la commission d'examen

Claude	Labit	Président
Michel	Barlaud	Rapporteurs
Anne	Guérin-Dugué	
Michel	Kerdranvat	Examinateurs
Henri	Sanson	
Patrick	Gros	Directeur de thèse
Christine	Guillemot	Co-directrice de thèse

Table des matières

Table des figures

Notations

Sauf mention explicite du contexte, les notations adoptées dans cette thèse sont les suivantes :

- les scalaires sont notées en minuscule, comme par exemple le scalaire x ;
- les vecteurs sont notées en minuscule et en gras, comme par exemple le vecteur $\mathbf{x}$;
- les matrices sont notées en majuscule, comme par exemple la matrice M ;
- ${}^tM$ est la matrice transposée de M ;
- $\overline{x}$ est le conjugué du complexe x ;
- $f(x, y)$ est la valeur au point (x, y) du signal continu f défini sur $\mathbb{R}^2$;
- $f[x, y]$ est la valeur au point (x, y) du signal discret f défini sur $\mathbb{Z}^2$;
- $L_2(\mathbb{R})$ est l'espace de Lebesgue des fonctions définis sur $\mathbb{R}$ à carré intégrable ;
- les fonctions régulières de classe C^p sont les fonctions p fois dérivables dont la dérivée d'ordre p est continue ;
- $\mathcal{M}_n(\mathbb{A})$ est l'ensemble des matrices carrés de taille $n \times n$ à valeur dans $\mathbb{A}$
- $[\![a, b]\!]$ est l'ensemble des entiers entre a et b, a et b inclus ;
- $|A|$ est le cardinal de l'ensemble A ;
- i est le complexe tel que $i^2 = -1$.

Par convention, un vecteur est un vecteur colonne.

Introduction

Le dernier standard JPEG 2000 a marqué un tournant dans l'évolution des schémas de compression d'images fixes. Le principal attrait du standard n'est pas apparu dans ses performances en compression, meilleures de 10 à 25% par rapport à celles du précédent standard JPEG [Chi03], mais dans les applications rendues possibles par une structuration plus élaborée des données. L'accroissement des espaces de stockage et des bandes passantes diminue considérablement la pression pour de meilleurs taux de compression. La demande est en revanche très forte pour élargir les normes de compression à l'ensemble des problèmes apparaissant dans une chaîne de communication d'images, comme la sécurité, la distribution, l'édition, la restauration, l'indexation et l'analyse d'images. JPEG 2000 est la première norme prenant en compte certains de ces aspects.

La partie 6 du protocole définit la manière de créer un document par la superposition d'un ensemble d'images, pouvant être naturelles, synthétiques, ou ne comporter que du texte, et d'adapter l'encodage à la spécificité de chacune des images. Cela permet d'optimiser le débit, d'éditer les images, et de disposer de l'information textuelle pour des applications en indexation et classification de documents. La partie 8 traite des problèmes de sécurité en définissant des protocoles de cryptage, de protection des droits de propriété, et d'accès à la résolution et aux parties de l'image en fonction des droits du client. La partie 9 est un protocole client-serveur permettant de distribuer le débit et les metadonnées adaptées au client, et de rendre prioritaire les régions d'intérêt définies par le client. Les industriels ont porté une certaine attention sur ces trois parties du standard. Il est néanmoins vraisemblable que ces nouvelles possibilités soient insuffisantes pour que le standard émerge. En particulier, un verrou majeur réside dans l'actuelle impossibilité d'interagir avec le client pour lui proposer de nouvelles images en vue de satisfaire sa demande. Le rapprochement entre des requêtes textuelles et des images non annotées, ou la compréhension du type de similarité existant entre différentes images, sont nécessaires pour élaborer des chaînes de communication interactive. Une telle avancée ne peut avoir lieu que si le standard permet un accès rapide et fiable à l'information visuelle caractérisant la scène ou le contenu. Cette thèse est, à travers une application concrète qu'est la détection de copies, une contribution dans cette voie.

Description et compression des images fixes

Jusqu'à une époque récente, la description et la compression étaient deux disciplines indépendantes. La pression pour l'émergence d'un standard de compression riche d'applications visuelles tend à faire disparaître cette frontière. Cette section introduit rapidement les problématiques de ces deux disciplines.

La vision est le processus d'extraction de caractéristiques 3D intrinsèques à la scène à partir d'une ou de plusieurs images. L'être humain est extrêmement performant dans la réalisation de cette tâche. Dans un laps de temps très court, les objets de la scène sont identifiés, et leur géométrie 3D est reconstruite malgré de sévères occlusions. La vision biologique est devenue un sujet d'étude scientifique au début du vingtième siècle avec les travaux de la Gestalt montrant que la compréhension d'une scène n'est pas obtenue par décomposition du stimulus visuel en entités indivisibles pré-définies. Les relations (de continuité, de proximité, de symétrie) existant entre différentes parties du stimulus créent un contexte porteur de sens. Leur détection est essentielle pour caractériser la scène observée. Sur le plan théorique, David Marr a énoncé au début des années 1980 un paradigme décomposant la vision en de nombreux modules distincts, comme la détection de contours, l'inférence de géométrie 3D à partir des contours, ou la reconnaissance de motifs répétés (texture). Ces modules permettent de représenter hiérarchiquement l'image, de la représentation en niveaux de couleur jusqu'à la description de la forme 3D de chacun des objets composant la scène. Les tâches réalisées par ces modules, la manière dont chacune d'entre elles est exécutée, et la façon dont est représentée la hiérarchie de description, restent pour la plupart des énigmes. L'extraction de caractéristiques par un module de vision, comme les contours ou la géométrie 3D inférée à partir de ces contours, n'est pas suffisante pour apprendre des modèles ou reconnaître des objets.

Pour identifier un objet ou pour guider le déroulement hiérarchique du processus de vision, le paradigme de Marr énonce également la nécessité de confronter les caractéristiques extraites avec celles de modèles connus. Lorsque l'espace des caractéristiques est un espace vectoriel normé et que la confrontation s'effectue au moyen d'une distance, la *description* désigne l'opération de vision réunissant les étapes d'extraction et de mesure de similarité. Au sommet du processus de vision biologique, les descripteurs sont dits de *haut niveau* d'abstraction, car ils portent la sémantique de la scène, ou encore le « sens »de la scène tel que perçu par un humain. En vision artificielle, les descripteurs ne sont aujourd'hui pas capables d'extraire une telle information, et sont dits de *bas niveau*. Le degré d'abstraction n'est, par conséquent, pas pris en compte dans l'évaluation des descripteurs. Leur performance est évaluée en mesurant leur *pouvoir discriminant* et leur *invariance* (ou *robustesse*). Le pouvoir discriminant et l'invariance désignent respectivement les facultés à séparer des stimuli de scènes différentes et à reconnaître une même scène prise dans des conditions variables de pose. Les descripteurs utilisés dans cette thèse sont *locaux*, c'est-à-dire calculés à partir d'une portion de l'image, par opposition aux descripteurs *globaux*. Les descripteurs globaux sont de très bas niveau, ils ne peuvent pas identifier les divers éléments constitutifs de la scène. Ils sont principalement

utilisés détection de « cuts »en vidéo, et en reconnaissance de texture en images fixes. La description locale, consistant à d'abord extraire des régions d'intérêt, puis à décrire indépendamment chacune d'entre elles, est beaucoup plus riche d'applications possibles. L'application retenue dans cette thèse est la détection de copies. La prochaine section introduit le problème de détection automatique de copies par le contenu, et la raison de ce choix. La section suivante présente les applications classiques de la description locale.

La compression s'occupe de trouver la quantité minimale d'information, ou *débit minimal*, permettant de reconstruire l'image originale à distorsion fixée. Il s'agit du problème essentiel de la communication d'images sur support contraint en bande-passante. Une image en niveaux de gris de résolution 1024×1024 quantifiée sur 8 bits requiert 1 Mo de bande-passante. Le problème de compression remonte aux années 1960 avec le premier système analogique de visio-conférence sur ligne téléphonique. Il requérait une large bande-passante pour la transmission d'une vidéo en noir et blanc et de faible résolution. Les progrès ont depuis été spectaculaires ; 0.1 bit par pixel suffit aujourd'hui pour obtenir une image de qualité moyenne. Même à reconstruction parfaite, il est possible d'obtenir des gains de débit importants. Cela est rendu possible par la forte dépendance statistique existant entre pixels proches. L'exploitation de dépendances simples comme la corrélation entre pixels voisins permet aux codeurs prédictifs ou aux transformées décorrélantes à coefficients entiers d'obtenir des taux intéressants de compression sans perte. La compression avec perte permet des gains bien plus importants, et est beaucoup plus répandue. Le système visuel humain est en effet insensible à de nombreuses pertes. Le schéma classique d'une compression avec perte s'effectue en trois étapes, composées d'une transformée d'image ayant pour propriétés d'être décorrélante, parcimonieuse ou compacte ; d'une étape de quantification visant à minimiser une mesure de distorsion pour un débit fixé ; et d'une étape de codage des coefficients quantifiés.

Dans les cas où la capacité de stockage ou de transmission impose aux images d'être compressées, la description devient très coûteuse : les images doivent être décompressées, puis transformées dans un espace adéquat, avant d'être finalement décrites. Et, ce, d'autant plus que les techniques de compression ont considérablement progressé, mais au prix d'une décompression de forte complexité. Une décompression JPEG requiert entre 150 et 300 instructions par pixel [SR96], soit plus de 40 millions d'instructions pour une image de résolution 512×512 (encore plus dans le cas de JPEG 2000). Il est donc intéressant de se poser dès aujourd'hui la question des conditions d'émergence d'un standard de compression permettant la réalisation de traitements visuels. Sur le plan scientifique, cette question a le mérite de rapprocher deux disciplines historiquement indépendantes, pouvant conduire à une réflexion nouvelle sur chacune d'entre elles. Les collaborations possibles entre compression et description sont nombreuses :

- investigation des représentations d'images utilisées en compression pour la description ;
- recherche de nouvelles caractéristiques visuelles compactes et discriminantes à partir de l'information minimale utilisée en compression pour reconstruire les images ;
- reconstruction des images à l'aide de modèles visuels définis par leur description.

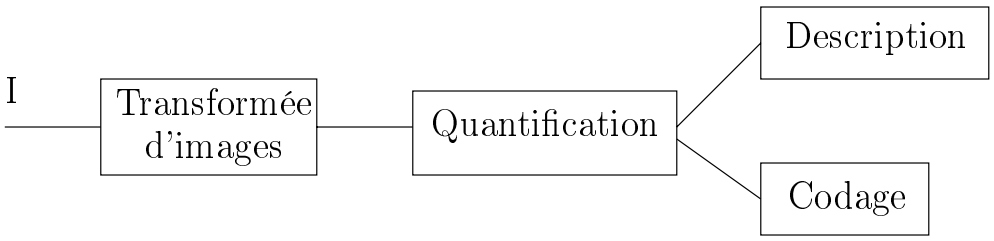

FIG. 1 – Schéma pour la compression et la description simultanées.

La collaboration visée dans cette thèse consiste à élaborer des schémas de compression permettant la description locale directement dans le domaine compressé ou pendant la reconstruction des images. La figure 1 illustre le schéma générique visé dans cette thèse. Le problème d'effectuer des tâches visuelles directement dans le domaine compressé est peu abordé dans la littérature. La première raison est d'ordre pratique. Les codeurs actuels empêchent la réalisation de toute tâche visuelle, ce qui rend nécessaire l'inversion de l'étape de codage. L'intérêt à pouvoir décrire à partir des coefficients quantifiés est limité par le fait que l'étape de décodage est la plus coûteuse dans les schémas de décompression. La seconde raison est d'ordre théorique. Le problème de connaître la quantité minimale d'information permettant d'atteindre une performance de description fixée n'a jamais été traité. Dans la théorie classique de l'information formulée par Shannon, la quantité minimale d'information peut se mesurer pour une distorsion fixée. Le choix de la mesure de distorsion s'effectue avant la minimisation, et détermine en retour la manière dont est altérée l'information par la compression. Lorsque l'image reconstruite est destinée à un humain, il est souhaitable que la distorsion mesure la qualité visuelle de l'image reconstruite. Une telle mesure est subjective, et le choix, largement répandu, de l'erreur quadratique moyenne s'avère adéquat. Ce n'est malheureusement plus le cas lorsque l'image compressée est également destinée à être décrite. Dans ce cas, l'information pertinente pour la description peut être très altérée à erreur quadratique moyenne faible. Il existe donc une difficulté dans l'élaboration théorique et la validation expérimentale d'un schéma de compression et de description simultanées.

Détection de copies par le contenu

Une copie est une image ayant subi une transformation. Dans la chaîne de communication d'images, certaines transformations sont très répandues comme la compression JPEG, ou la création de vignettes à résolution moindre. Les transformations les plus difficiles à détecter sont celles dont l'objectif est de camoufler un acte de piraterie. L'ensemble des transformations à disposition du pirate est réduit dans cette thèse à l'ensemble des *transformations admissibles*, composé des transformations monotones de luminance, des similitudes (groupe des translations, des rotations, et des changements d'échelle), et des coupures (ou *crops* en anglais). L'ensemble des transformations admis-

sibles est en fait beaucoup plus large, mais l'expérience montre qu'une détection robuste à ces quelques transformations est en fait robuste à un ensemble beaucoup plus vaste de transformations. Si le signal image est fortement altéré par une transformation admissible, un humain reconnaît toutefois aisément qu'il s'agit d'une copie, car les deux scènes sont identiques. La détection de copies peut donc être considérée comme un problème de vision, où *l'information utile* à extraire sur l'image doit permettre de partitionner l'ensemble des images en deux classes: l'ensemble des images issues de cette scène, et l'ensemble de toutes les autres images. Cette partition peut s'effectuer par extraction de descripteurs sur l'image requête, et par mesure de similarité de ces descripteurs avec ceux de la base. La détection de copies peut donc être considérée comme une opération de description d'images. Dans un tel schéma, une image n'étant pas une copie, mais provenant d'une même scène qu'une image de la base, conduira vraisemblablement à une fausse alarme. Il est en théorie possible de réduire les fausses alarmes aux images d'une même scène prises sur un même axe optique. La détection de copies étant dans cette thèse plus considérée comme un protocole expérimental permettant de mesurer la performance des descripteurs, ce travail de réduction des fausses alarmes n'est pas traité.

Le problème de détection de copies est récent et n'était pas considéré comme un problème de vision à l'origine, mais comme un problème de tatouage. Avant de mettre en ligne une image, et donc de l'exposer à la piraterie, le tatouage consiste à insérer une marque invisible dans l'image dont le but est d'authentifier le propriétaire de l'image. En plus d'être invisible, la marque doit être robuste aux transformations admissibles, et l'algorithme d'insertion de marque ou les éventuelles clés privées utilisées dans ces algorithmes doivent être indétectables. Tout cela rend le problème très difficile. Malgré une décennie de recherche, les algorithmes de tatouage peinent à être utilisés en pratique. Il en existe trois types. Le tatouage privé nécessite l'image originale lors de l'extraction de marque, ce qui n'est pas viable pour des bases d'images de volume important. Le tatouage aveugle requiert également un tiers de confiance dépositaire d'un secret. Ce secret n'est pas l'image originale, mais la clé de chiffrement utilisée lors de l'insertion de marque. Les deux principaux obstacles au tatouage aveugle sont d'une part la lourdeur de l'infrastructure à mettre en oeuvre pour disposer d'un tiers de confiance, et d'autre part la difficulté à obtenir la robustesse aux transformations géométriques. Dans le cas des rotations par exemple, il faut soit disposer d'une insertion de marque invariante aux rotations, soit tester toutes les rotations lors de l'extraction de marque. En pratique, les techniques reposant sur une insertion invariante sont moins robustes. Les algorithmes les plus performants requièrent donc de tester toutes les transformations géométriques, rendant l'étape d'extraction de marque très coûteuse en temps de calcul (supérieur à la minute). Il existe en théorie un troisième type de tatouage, dit asymétrique, qui ne requiert aucun tiers de confiance. Toutefois, il n'existe pas d'algorithme valable de marquage asymétrique et on doute de jamais en trouver. Précisons enfin que la détection d'une marque n'est pas en une preuve juridique de copie. L'ensemble des techniques de tatouage et leurs limitations peut se trouver dans [PAK99].

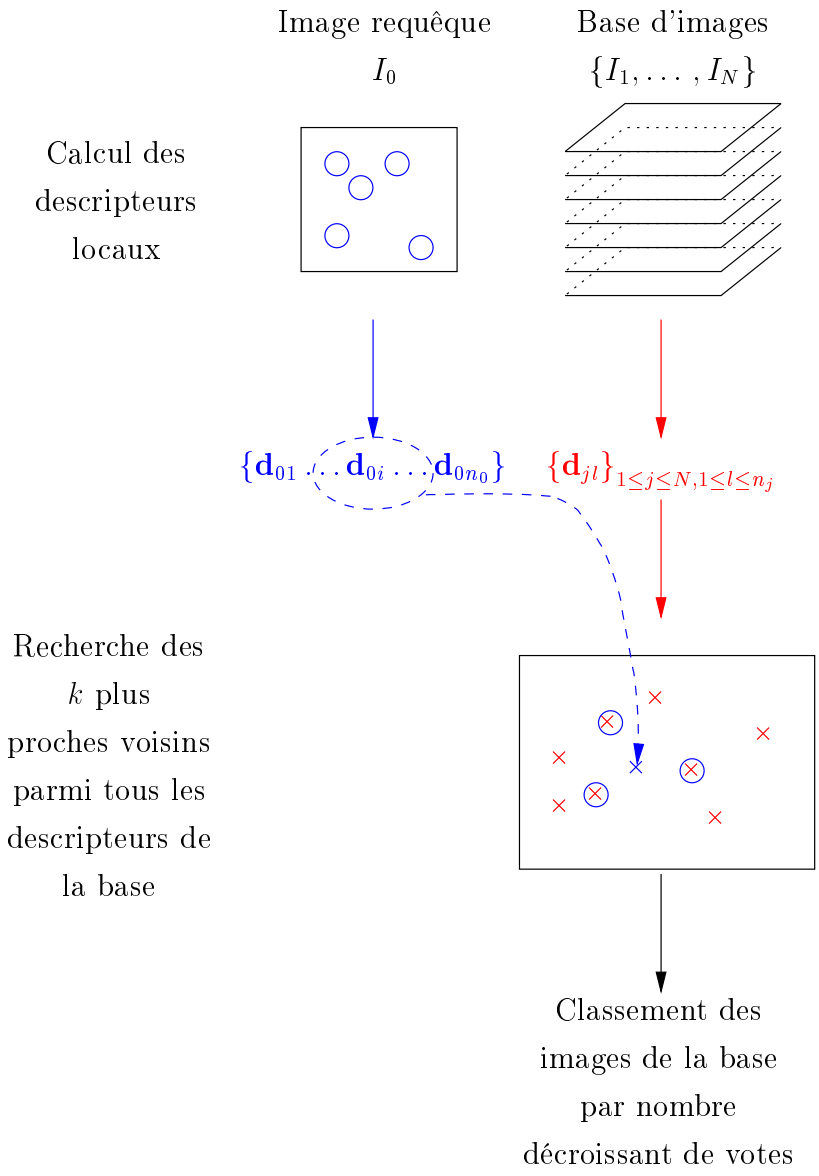

FIG. 2 – Schéma de recherche d'images par le contenu décomposé en étapes de description et de votes.

Le principe de détection de copies par le contenu est simple : une similarité anormalement élevée entre les descripteurs d'une image requête et ceux d'une image de la base à protéger conduit à un soupçon de copie. Les descripteurs utilisés dans cette thèse sont locaux. Le seul problème de détection de copies est sans doute mieux traité par l'utilisation des deux types de description, locale et globale. Toutefois, les applications de la description globale étant réduites, seule la description locale est utilisée dans cette thèse. La détection de copies sert de protocole expérimental pour la mesure de la performance des descripteurs. La description locale consiste à d'abord extraire les régions d'intérêt présentes dans l'image, puis à décrire indépendamment chacune d'entre elles. Dans la littérature est rarement abordé le problème d'extraction de caractéristiques de plus haut niveau à partir de la spécificité de l'agencement spatial entre les régions décrites. La difficulté principale est le coût algorithmique d'une telle extraction et la robustesse des caractéristiques ainsi extraites. Dans les schémas de bas niveau existants, une image est représentée par un ensemble de points dans l'espace de description. Pour limiter la complexité de la mesure de similarité entre nuages de points appartenant à un espace de dimension élevée (typiquement entre 10 et 150), le schéma de détection de copies adopté dans cette thèse est celui classiquement utilisé dans les systèmes de recherche d'images par le contenu [BAG03, SM97]. Ce schéma, représenté par la figure 2, se déroule en trois étapes. L'image requête est l'image pour laquelle le système doit décider s'il s'agit d'une copie d'une des N images de la base $\{I_j\}_{1\leq j\leq N}$. Chaque image I_j est décrite hors ligne par n_j descripteurs locaux $\{\mathbf{d}_{jl}\}_{1\leq l\leq n_j}$. La détection en ligne consiste en les trois étapes décrites ci-après.

1. Calcul des n_0 descripteurs $\{\mathbf{d}_{0i}\}_{1\leq i\leq n_0}$ de l'image requête,
2. Recherche, pour chaque descripteur requête $\mathbf{d}_{0i}$, des k descripteurs de la base les plus proches au sens d'une norme fixée, en général la distance euclidienne. Le nombre k de plus proches voisins est fixé, et typiquement situé entre 10 et 100 pour 10^8 descripteurs dans la base. Un vote est accordé à chacune des images de la base ayant un de leurs descripteurs parmi les k plus proches du descripteur requête.
3. Classement des images de la base par nombre décroissant de votes.

Des travaux récents [Lej05] en indexation de grandes bases d'images permettent d'effectuer une recherche approximative de plus proches voisins en un temps largement sous-linéaire, de 100 à 1000 fois inférieur à celui d'une recherche séquentielle. Le temps de détection de copies est désormais faible, de l'ordre de cinq secondes pour 10^5 images ($\simeq 10^8$ descripteurs), c'est-à-dire, pour cette taille de base, largement inférieur au temps de détection de marque, qui est de l'ordre de la minute.

Objectifs et approches

Cette thèse vise à élaborer des schémas de compression adaptés à la description locale. Ce problème est peu abordé dans la littérature. Les travaux existants connexes

à ce problème peuvent se classer en trois catégories.

La première catégorie porte sur la recherche de représentations d'images adaptées simultanément à la compression et à la description. Les représentations permettant d'atteindre de bons taux de compression, c'est-à-dire d'entropie et de redondance faibles, sont désormais très nombreuses. Il est intéressant de constater leur grande variété: représentations linéaires ou non, isotropes ou directionnelles, adaptatives ou fixes. Pour le problème de description ou pour d'autres traitements visuels, les représentations doivent être équivariantes au groupe des similitudes, composées des translations, des rotations et des changements d'échelle [SAH92, Kin98]. Le choix de la représentation d'images est essentiel et détermine en grande partie la performance qu'il est possible d'atteindre en compression et description.

La seconde catégorie de travaux concerne la *description globale* à partir de représentations appropriées à la compression [dWSD99, DV02]. L'information contenue dans un descripteur global porte sur toute l'image, et ne permet donc pas l'identification des divers éléments constitutifs de la scène. Ils sont donc de bas niveau, et généralement d'intérêt applicatif plus faible que les *descripteurs locaux*, dont l'information localisée permet une meilleure compréhension de la scène. De plus, dans les travaux existants, les descripteurs sont calculés à partir de la représentation non quantifiée. Le problème consistant à compresser dans le but de préserver l'information pertinente pour la description reste ouvert, même dans le cas d'une description globale. Le problème de description à partir de représentations compressées a plus été traité pour les vidéos que pour les images fixes. L'information de compensation de mouvement présente dans le standard MPEG est portée par des vecteurs pointant sur des blocs similaires. Selon les caractéristiques du bloc, les vecteurs de compensation peuvent être des vecteurs de mouvement au sens du flot optique, et donc être utilisés pour de nombreuses tâches, comme la détection automatique de « cuts », la caractérisation du mouvement de la caméra, l'indexation par la description du mouvement [CB05].

La troisième catégorie rassemble un nombre limité de travaux portant sur l'extraction robuste de points d'intérêt à partir de représentations en ondelettes [LS99, CLS95]. Ces représentations ont permis des progrès considérables en compression d'images, mais là encore, l'extraction de points, qui constitue la première étape du processus de description locale, est effectuée à partir des représentations non quantifiées.

La description locale dans le domaine compressé est riche d'applications. Les principales sont décrites ci-dessous.

La mise en correspondance de plusieurs images d'une même scène, dites homologues. Ce problème consiste à établir une correspondance point à point entre les images. La contrainte épipolaire réduit la recherche du correspondant aux points localisés sur une droite de l'image homologue. Dans le cas de caméras non étalonnées, cette contrainte n'est pas connue et l'ensemble des correspondances possibles entre points est extrêmement grande. Pour réduire cet ensemble, un faible nombre de points est extrait sur chacune des images. Le voisinage de chacun des points extraits est décrit. Les points

sont ensuite appariés par une simple recherche du descripteur le plus proche. Enfin, des contraintes géométriques sont utilisées pour éliminer les mauvais appariements. Les techniques de description sont donc très utiles pour initialiser la mise en correspondance complète à partir d'un nombre réduit de points. Une application de l'appariement est la création de mosaïques d'images directement sur l'appareil numérique, permettant à l'utilisateur de visualiser in situ la composition. Une seconde application est l'étalonnage de caméras en stéréo-vision. Lorsque les caméras transmettent leurs images compressées (à cause d'une liaison bas débit, typiquement sans fil), la capacité de l'unité réceptrice à étalonner directement dans le domaine compressé peut être avantageuse.

Le suivi de points ou d'objets à partir d'images séparées dans le temps est similaire au cas précédent. L'appariement de points d'intérêt par la recherche de descripteurs les plus proches permet d'estimer le mouvement en un nombre réduit de points et donc d'initialiser le suivi des objets d'intérêt. Décrire dans l'espace compressé permettrait à l'unité réceptrice de piloter automatiquement et en temps réel une caméra sans fil. Dans cette thèse, les schémas de compression pour la description locale sont conçus pour les images fixes. Le problème de l'extension de ces schémas de compression à la vidéo n'est pas traité et reste ouvert. Précisons toutefois que les représentations d'images utilisées dans cette thèse sont équivariantes aux similitudes, donc propices à l'estimation de mouvement.

La détection et la reconnaissance d'objets. Ce sont des problèmes difficiles car ils supposent d'identifier un objet à la fois dans des conditions changeantes de prises de vue et dans les aspects variables d'existence de cet objet. Les méthodes de reconnaissance utilisant la description locale [Low99, BMP02] figurent parmi les plus performantes. Pour chaque objet à reconnaître, un modèle est construit lors d'une phase d'apprentissage. Le modèle consiste en l'ensemble des descripteurs calculés à partir d'images homogènes représentant l'objet seul sous différentes prises de vue. Lors de la phase de reconnaissance sur une image hétérogène, les descripteurs sont calculés, et une technique soit probabiliste, soit par clustering, est mise en oeuvre pour détecter le nombre d'objets présents dans l'image. Les descripteurs ont beaucoup évolué en reconnaissance d'objets. Ils ont d'abord été des descripteurs de forme, calculés à partir des contours de l'image segmentée. Ils ont ensuite été les descripteurs locaux présentés précédemment, c'est-à-dire des descripteurs d'apparence. La tendance actuelle est dans la conception de descripteurs prenant compte de la forme et de l'apparence. Reconnaître dans l'espace compressé présentera un intérêt lorsque le temps de reconnaissance sera plus faible ou de l'ordre du temps de décompression.

La recherche d'images par le contenu. Très souvent, la sélection des images d'une grande base susceptibles d'intéresser un utilisateur est effectuée au moyen de mots-clés. La reconnaissance d'objets n'est pas encore assez avancée pour permettre l'annotation automatique des images. La description locale, introduite dans [SM96], a

ouvert une voie prometteuse dans cet axe de recherche à long terme. Contrairement à la description globale, ce type de description permet en effet la détection et la reconnaissance des différents éléments constitutifs de la scène. L'évaluation des descripteurs locaux peut s'effectuer par leur capacité à retrouver les images de la base présentant une similarité visuelle avec une image requête. Pour s'affranchir de la subjectivité de la mesure de similarité, une solution consiste à disposer dans la base de plusieurs images d'une même scène prise dans des conditions variables de pose et d'éclairage. Une autre solution consiste à choisir comme image requête une image de la base ayant subi une transformation synthétique, comme une rotation, un changement d'échelle, un crop, ou un filtrage (linéaire ou non). Cette dernière solution présente l'intérêt d'avoir la détection de copies comme application directe. L'évolution rapide des techniques de description et la sécurité du système de détection de copies nécessitent de recalculer les descripteurs régulièrement. Un temps considérable peut être économisé en effectuant directement la détection dans le domaine compressé.

Aucun schéma complet permettant la description locale dans le domaine compressé n'a, à ce jour, été proposé. Pour montrer l'existence de tels schémas, l'application choisie dans cette thèse est la détection de copies. Outre son intérêt pratique, cette application constitue un bon cadre d'évaluation. Les limites de performance peuvent être approchées par le volume de la base d'images à protéger et par la nature des transformations synthétiques subies par les copies. Les sources d'inspiration permettant d'élaborer de tels schémas se trouvent dans l'existant en description et en compression. Les techniques de description seront revisitées avec une attention portée sur la quantité minimale d'information qu'elles requièrent, en particulier sur la redondance des représentations d'images utilisées. De nombreuses représentations sont apparues et ont contribué de façon significative à l'amélioration des techniques de compression. Ces représentations seront analysées en vue d'évaluer leur propension à élaborer des schémas de description locale. Ces deux analyses préalables permettront de retenir les pyramides laplaciennes et les représentations multirésolution orientables comme celles permettant le meilleur compromis entre redondance et stabilité. Finalement, une méthode sera proposée pour compresser ces deux types de représentations tout en préservant l'information requise pour la description locale. Le cadre d'évaluation du schéma final consiste en la détection de copies à partir d'une base de 30 000 images compressées à différents débits.

Contributions et contenu de la thèse

Le plan de la thèse suit l'approche proposée dans la section précédente. Cette section décrit plus en détail chacun des chapitres et leurs principales contributions.

Le chapitre 1 présente un état de l'art partiel des techniques de description locale. Dans un premier temps sont caractérisées les représentations d'images adaptées à ce problème. Le choix arbitraire de la grille sur laquelle est définie l'image en niveaux de gris est le premier obstacle à l'extraction de caractéristiques intrinsèques à la scène.

Il est montré que les représentations covariantes aux similitudes permettent en partie de s'affranchir de ce choix. Cette covariance impose aux représentations d'être multi-échelles et isotropes, ou multi-échelles et multi-orientations. Dans un second temps sont présentées les principales techniques de description locale. Elles sont composées d'une étape d'extraction robuste de points localisés dans un espace paramétré en position, en échelle et éventuellement en orientation, suivie d'une étape de description des voisinages de chacun des points extraits. La troisième partie introduit le protocole expérimental utilisé dans cette thèse pour évaluer la performance de description.

Le chapitre 2 examine certaines techniques classiques utilisées en détection robuste de points, d'échelles, et d'orientations. Ces attributs sont nécessaires pour le calcul du descripteur SIFT [Low99] faisant aujourd'hui référence pour sa robustesse et son pouvoir discriminant. L'extraction d'échelle caractéristique proposée dans [Lin94b] et les extracteurs de points introduits dans [Lin94b, SM96] sont analysés sous un angle neuf consistant à évaluer l'influence de la discrétisation du paramètre d'échelle sur leur robustesse. La finesse de la discrétisation détermine la redondance de la représentation d'images ; cette évaluation est donc nécessaire pour la recherche de la quantité minimale d'information pour la description. De même, pour l'étape de description, la technique de référence proposée dans [Low99] est analysée en détail. Ce descripteur, dénommé SIFT, repose sur les distributions locales des orientations, estimées à partir du gradient. La détection robuste d'orientation a fait l'objet de nombreux travaux depuis ceux de Knutsson [GK95]. D'une part est évalué le gain en robustesse d'estimation de l'orientation en utilisant certaines de ces méthodes par rapport à celle du gradient utilisé dans SIFT. D'autre part, le pouvoir discriminant du descripteur SIFT est évalué en fonction de la discrétisation en orientation. Le but principal de ce chapitre est d'évaluer la redondance minimale des représentations utilisées en description.

Le chapitre 3 présente brièvement les représentations d'images utilisées en compression. La seule alternative aux représentations en niveaux de gris a longtemps été la représentation de Fourier et celle de Gabor. Il en existe aujourd'hui beaucoup d'autres, comme les représentations en ondelettes, les représentations orientables, les pyramides laplaciennes, les décompositions par « matching pursuit ». Leur apparition et leur évolution sont en grande partie liées à la recherche de représentations adaptées aux images naturelles, c'est-à-dire pour lesquelles l'énergie se concentre en un nombre réduit de coefficients. Ceci est rendu possible par la construction de dictionnaires dont les éléments sont des structures génériques des images naturelles, comme des contours par exemple. Un tel dictionnaire offre la possibilité de décrire les images en termes d'éléments caractéristiques, plus pertinents que la valeur initiale de luminance. La description n'est néanmoins possible que si la distribution de l'énergie sur le dictionnaire est stable. Si la représentation d'une image faiblement perturbée est très différente, aucune description n'est possible. Ce critère de stabilité est la source de l'antagonisme présent dans la recherche de représentations adaptées simultanément à la compression et à la description. Les représentations multi-échelles utilisées en compression sont efficacement implémentées par des bancs de filtres partageant de nombreuses caractéristiques, comme

l'échantillonnage en espace et en échelle, et le spectre, la symétrie, la régularité, ou l'orthogonalité des filtres utilisés. Une attention est donc portée sur le lien entre ces caractéristiques et la représentation d'images, et plus particulièrement sur les causes de la variance aux translations des représentations en ondelettes utilisées dans le standard JPEG 2000. Malgré cette variance, une méthode inspirée de [LS99] est proposée pour extraire des points d'intérêt. Une tentative est également faite pour limiter la variance aux rotations, mais les résultats sont décevants. La conclusion de ce chapitre est que les transformées à échantillonnage critique sont trop instables et donc inadaptées pour la description. La très bonne robustesse des points extraits à partir des représentations non sous-échantillonnées laisse toutefois espérer qu'une faible redondance est suffisante pour obtenir une description stable.

Le chapitre 4 présente des représentations stables et redondantes. Les représentations en ondelettes isotropes et non sous-échantillonnées sont trop redondantes pour une compression efficace. Il est néanmoins important d'évaluer la robustesse des points extraits à partir de celles-ci, en vue de valider l'échantillonnage en échelle classiquement utilisé en compression. Cet échantillonnage est en effet beaucoup plus grossier que celui utilisé en description, il n'est que d'une voie par octave. Les tests montrent que, sans échantillonnage spatial, un tel échantillonnage en échelle est suffisant pour la description. Ensuite sont étudiées les représentations par pyramide laplacienne [BA83], qui sont proches de représentations en ondelettes isotropes, mais échantillonnées en espace. Elles permettent donc d'évaluer la dégradation causée par le sous-échantillonnage. Cette dégradation est suffisamment faible pour obtenir des points d'intérêt et des descripteurs locaux stables. Toutefois, l'estimation d'orientation, nécessaire pour le calcul des descripteurs SIFT, impose d'inverser la pyramide laplacienne. Par la suite sont donc analysées des représentations anisotropes, pour lesquelles le schéma complet de description locale peut être exécutée dans le domaine transformé. Les contourlets [DV00] constituent l'extension naturelle de la pyramide laplacienne aux transformées anisotropes. La décomposition directionnelle à échantillonnage critique provoque une baisse sensible de performance. Un soin particulier doit être apporté dans le choix des filtres et la conception du banc de filtres, en vue de réduire le recouvrement spectral. La transformée en ondelettes complexes [Kin98] est, en ce sens, intéressante. Toutefois, les coefficients de cette transformée sont définis sur une grille dyadique dont la résolution la plus fine n'est, comme dans le cas des ondelette réelles, que le quart de la résolution de l'image initiale. Cette résolution n'est pas suffisante pour extraire des points d'intérêt de façon robuste. Les transformées orientables [SAH92] sont des candidates naturelles pour l'estimation robuste d'orientation. Ces transformées décomposent l'image en un nombre réduit de bandes orientées, et ont la particularité de permettre l'interpolation de la transformée en tout autre orientation. La performance de l'extraction de points et de la description est, dans ce type de représentations, proche de celle des références existantes.

Le chapitre 5 présente des schémas de compression adaptés à la description locale. Il s'agit d'une contribution de cette thèse. Le problème de savoir comment dégrader l'image pour gagner en débit et ne pas perdre en potentiel de description n'a en effet

jamais été abordé. Deux représentations fort différentes sont examinées : la pyramide laplacienne, isotrope et de faible redondance ; et les représentations orientables, directionnelles et de forte redondance. Dans les deux cas est évalué l'impact sur la description de la quantification et d'une technique de compression des représentations redondantes dénommée POCS (Projection Onto Convex Sets). Les résultats montrent la meilleure performance des pyramides laplaciennes sur les représentations orientables en terme de compromis entre PSNR, entropie, et répétabilités. Toutefois, le coût de calcul de la description à partir des pyramides laplaciennes est plus élevé, car ce type de représentation isotrope requiert la reconstruction progressive de l'image pour estimer l'orientation. Le coût peut être beaucoup plus faible à partir des représentations orientables. Les techniques existantes de compression pour ce type particulier de représentations sont présentées ; puis une nouvelle technique de codage adapté à la description est proposée. Enfin, la validation de ces schémas de compression est effectuée par détection de copies à partir d'une base hétérogène de 30 000 images compressées à différents débits.

Chapitre 1

Description locale dans l'espace-échelle gaussien

Le problème conjoint de compression et de description locale des images ne peut être résolu que s'il existe une représentation adaptée, c'est-à-dire permettant simultanément de reconstruire correctement l'image à partir d'une quantité minimale d'information et d'acquérir une information pertinente sur la scène. Ce chapitre vise à introduire les méthodes couramment utilisées en description. La première section restreint les représentations adaptées au problème de description comme étant celles qui sont linéaires et covariantes aux similitudes du plan. Ces représentations sont continûment paramétrées en échelle, et isotropes ou continûment paramétrées en orientation. Deux représentations naturellement candidates sont l'espace-échelle gaussien et les représentations continues en ondelettes. La seconde section parcourt rapidement les techniques existantes en description locale, et introduit les protocoles qui seront utilisés dans cette thèse pour évaluer la performance de cette description. Les descripteurs actuels les plus performants requièrent qu'une échelle et une orientation soient détectées en tout point d'intérêt. La troisième section évalue la précision de détection de ces paramètres d'échelle et d'orientation en fonction de la discrétisation de la représentation. Cette section évalue également la performance d'un descripteur de référence, nommé SIFT, en fonction de cette même discrétisation en échelle et en orientation de la représentation. Ces deux évaluations permettent d'obtenir la redondance minimale des représentations pour une performance de description fixée. Cette redondance doit être suffisamment faible pour qu'une représentation adaptée au problème conjoint de compression et de description existe. Les performances de description obtenues avec une fine discrétisation en échelle et en orientation, c'est-à-dire avec une représentation de forte redondance, serviront à majorer les performances qu'il sera possible d'obtenir dans les prochains chapitres.

1.1 Représentations d'images pour la description locale

Dans cette thèse, les images seront souvent discrètes et quelquefois continues.

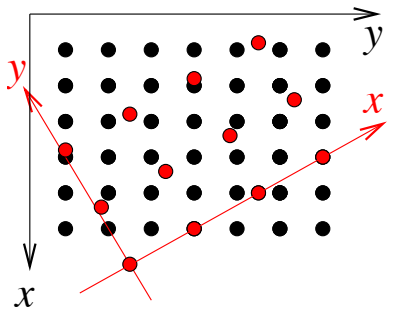

FIG. 1.1 – Dépendance d'une image vis à vis du choix de la grille.

Définition 1 *Une image discrète (resp. continue) en niveaux de gris est un élément de l'espace image* Ω*, ensemble des fonctions de* $S = \mathbb{N}^2$ *(resp.* $S = \mathbb{R}^2$*) dans* $V = [\![0, 255]\!]$ *(resp.* $V = [0, 1]$*) à support rectangulaire :*

$$\Omega = \{I : S \to V / \ \exists (x_o, y_o) \in S \ : \ \forall |x| > x_o, \forall |y| > y_o, \ I(x, y) = 0\}$$

L'ensemble des images discrètes nulles sur $\mathbb{N}^2 - [\![1, n]\!] \times [\![1, m]\!]$ *sera noté* Ω_{nm}

Le problème de vision consistant à retrouver des caractéristiques de la scène 3D à partir d'une image 2D est un problème inverse mal posé (non unicité de la solution) et difficile (l'information de luminance échantillonnée sur une grille est loin de l'information utile cherchée). Une première difficulté apparaît avec la nécessité de dériver les caractéristiques de la scène 3D indépendamment de la grille sur laquelle est définie l'image. Les grilles de la figure 1.1 diffèrent par une translation et une rotation de leurs repères, et par leur résolution. Une tâche élémentaire de tout système de vision, et en particulier de tout système de description, est de reconnaître que deux images formées sur différentes grilles proviennent d'une même scène. La première étape du processus de description consiste donc à transformer l'espace image en un espace d'observation permettant de s'affranchir au mieux du choix arbitraire de la grille. L'objet de cette section est de caractériser cette transformation.

1.1.1 Contraintes de linéarité

Seules les transformations linéaires seront considérées dans cette thèse. Ce choix ne se justifie que par l'ampleur de l'investigation qu'il reste à mener pour les transformations linéaires. Les fréquentes occultations apparaissant dans les images naturelles incitent à penser que les transformations non linéaires sont mieux adaptées pour extraire l'information visuelle. Des travaux ont déjà été menés en ce sens, et s'intensifient aujourd'hui [Mal91, HS05, JSIN06]. Les traitements non linéaires semblent indispensables dans une chaîne de traitement visuel ; toutefois, ils peuvent apparaître après une première transformation linéaire. Cette première transformation a pour but de réduire la redondance d'information dans le stimulus visuel, et de permettre un accès rapide et fiable à l'information visuelle pertinente.

Se donnant une famille $\{\tilde{\varphi}_k\}_{k\in K}$ génératrice de Ω_{nm}, toute image I de taille $N = n \times m$ se décompose en une combinaison linéaire des *fonctions de base* :

$$I = \sum_{k\in K} c_k \tilde{\varphi}_k \tag{1.1}$$

où les coefficients transformés c_k s'obtiennent par projection de l'image I sur une famille de *fonctions de projection* $\{\varphi_k \in \Omega_{nm}\}_{k\in K}$:

$$c_k = \langle I|\varphi_k\rangle = \sum_{n_1,n_2\in \mathbb{N}} I(n_1,n_2)\overline{\varphi}_k(n_1,n_2) \tag{1.2}$$

L'ensemble de ces coefficients transformés constitue la *représentation de l'image* I dans la famille $\{\varphi_k\}_{k\in K}$. La transformation associée à cette représentation est l'application T définie sur Ω_{nm} qui à une image I associe la fonction réelle ou complexe définie sur K par $T[I](k) = \langle I|\varphi_k\rangle$. La relation 1.1 montre que l'application T est inversible à gauche. En pratique, K est fini de cardinal $M \geq N$. Si l'on écrit l'image I sous sa forme vectorisée $\mathbf{x} \in \mathbb{R}^N$, la relation 1.2 montre que la transformée vectorisée $\mathbf{y} \in \mathbb{R}^M$ peut s'écrire :

$$\mathbf{y} = T\mathbf{x} \tag{1.3}$$

où T est la matrice caractérisant la transformation. Cette matrice est égale à :

$$T = {}^t[\overline{\varphi}_1\overline{\varphi}_2\ldots\overline{\varphi}_M] \tag{1.4}$$

où $\overline{\varphi}_k$ est le vecteur colonne composé des $\overline{\varphi}_k(n_1,n_2)$, $(n_1,n_2) \in [\![1,n]\!] \times [\![1,m]\!]$. Si $M = N$, les familles $\{\overline{\varphi}_k\}_{1\leq k\leq N}$ et $\{\tilde{\varphi}_k\}_{1\leq k\leq N}$ sont libres et forment des bases respectives de l'espace transformé et de l'espace image Ω_{nm}. Si $M > N$, la transformation T est redondante et possède une infinité d'inverses à gauche dont le pseudo-inverse défini par $({}^tTT)^{-1}$. Les transformations redondantes sont de plus en plus étudiées en vision, comme en débruitage, en analyse de texture, ou même en compression. Les *frames* constituent le cadre de leur étude.

Définition 2 *La famille $\Phi = \{\varphi_k\}_{k\in K} \subset H$ ($(H, ||.||)$ espace vectoriel normé) est un frame de H s'il existe $A > 0$ et $B < \infty$ tels que :*

$$\forall \mathbf{x} \in H,\ A||\mathbf{x}||^2 \leq \sum_{k\in K} |\langle \mathbf{x}|\varphi_k\rangle|^2 \leq B||\mathbf{x}||^2 \tag{1.5}$$

A et B sont les bornes du frame.

L'existence de $A > 0$ assure que la famille Φ est génératrice de H. En dimension finie (égale à M), l'existence d'une borne supérieure B finie est toujours vraie puisqu'on peut choisir $B = \sum_{k=1}^{M} ||\varphi_k||^2$.

Définition 3 *La redondance d'un frame fini $\Phi = \{\varphi_k\}_{1\leq k\leq M}$ générateur de l'espace vectoriel H de dimension N est $r = \frac{M}{N}$. La redondance de la transformation T associée au frame Φ est donc égale au nombre de coefficients transformés sur le nombre de pixels de l'image originale.*

Lorsque $A = B$, le *frame* est dit ajusté, et si les éléments φ_k du frame sont normalisés, on peut choisir $A = r$. Lorsqu'un *frame* est ajusté et libre, la relation 1.5 devient la relation de Parseval ; le *frame* constitue dans ce cas une base orthonormée, et la transformation T associée est unitaire : ${}^tTT = Id$.

Une transformation adaptée au problème conjoint de compression et de description doit être de faible redondance et covariante aux similitudes. L'antagonisme entre ces deux critères apparaîtra fréquemment dans les prochains chapitres, et une problématique centrale réside dans la manière de construire un compromis intéressant entre ces deux critères.

1.1.2 Contraintes de covariance

En physique, pour effectuer des mesures sur un champ suffisamment différentiable de scalaires, le repère n'est pas choisi par l'observateur mais est fixé par le champ lui-même. En tout point est défini le repère orthonormal direct $(\mathbf{n_v}, \mathbf{n_w})$, où $\mathbf{n_w}$ pointe dans la direction du gradient. La figure 1.2 montre ce repère pour des champs 1D et 2D. Les mesures obtenues dans ce repère locale sont *invariantes* à toute transformation du repère sur lequel est défini le champ. Ce repère a déjà été utilisé pour construire des descripteurs, comme par exemple les invariants différentiels présentés dans la section 1.2.2. Ce repère local n'est en revanche d'aucune aide en compression, puisqu'un repère global est nécessaire pour définir le champ en tout point. Une autre façon de s'affranchir du choix du repère consiste à obtenir des mesures non pas invariantes mais *covariantes* aux changements de repère.

Définition 4 *La représentation $T[I]$ de l'image $I \in \Omega$ est dite covariante à la transformation géométrique $t : \Omega \to \Omega$ si $T[t \circ I] = t \circ T[I]$.*

Dans le cas de mesures covariantes aux changements de repère, les mesures commutent avec les transformations subies par le repère, et il est donc toujours possible de ramener les mesures dans un repère de référence. Pour cela, il est nécessaire que les mesures, c'est-à-dire les coefficients transformés c_k de la relation 1.2, soient indicées sur le même repère que l'image originale $I \in \Omega_{nm}$. Le support des transformations d'images $T[I]$ cherchées est de la forme $[\![1, n]\!] \times [\![1, m]\!] \times K$. Dans la suite sont caractérisées les transformations linéaires covariantes aux changements de repères orthogonaux, c'est-à-dire aux similitudes (groupe des translations, rotations, et homothéties). On verra qu'une telle covariance peut être obtenue en ajoutant une dimension dans l'espace transformé par degré de liberté dans le changement de repère orthogonal. S'affranchir du choix arbitraire de la grille par une transformation covariante est donc une solution très coûteuse en terme de volume de données, puisque l'espace transformé est beaucoup plus grand que l'espace initial. Elle est inappropriée au problème de compression, sauf s'il est possible de discrétiser suffisamment grossièrement les dimensions supplémentaires. C'est le compromis entre redondance et covariance que les prochains chapitres viseront à trouver.

FIG. 1.2 – Repère local invariant aux translations et rotations, pour un signal 1D et un signal 2D.

Covariance aux translations. L'observation doit être indépendante du choix du centre du repère sur l'image I. La transformation T est donc covariante aux translations :

$$T[I_{\mathbf{u}}](\mathbf{x}) = T[I](\mathbf{x}+\mathbf{u})$$

où $I_{\mathbf{u}}$ est l'image I exprimée dans un repère translaté de $\mathbf{u} = (u_1, u_2) \in \mathbb{Z}^2$. On parle alors de représentation *homogène* pour exprimer le fait que le même traitement est effectué en tout point de l'image. La linéarité et la covariance aux translations conduisent à une transformation T de la forme :

$$\begin{aligned} T[I] &= T[\sum_{p_1,p_2\in\mathbb{Z}} I(p_1,p_2)\delta(.-p_1,.-p_2)] \\ &= \sum_{p_1,p_2\in\mathbb{Z}} I(p_1,p_2)T[\delta(.-p_1,.-p_2)] \\ &= \sum_{p_1,p_2\in\mathbb{Z}} I(p_1,p_2)T[\delta](.-p_1,.-p_2) \\ &= I * T[\delta] \end{aligned}$$

où δ est l'image Dirac (noire partout sauf en un point, à l'origine, où elle est blanche). Une représentation linéaire covariante aux translations est donc une convolution entre l'image I et un filtre d'analyse $h = T[\delta]$. Lorsque la réponse impulsionnelle est finie, le filtre intègre la luminance sur une certaine zone, appelée champ réceptif en vision biologique, et de taille égal au support du filtre d'analyse. Pour le problème conjoint de compression et de description, les caractéristiques essentielles du filtre sont la parcimonie de représentation des images naturelles, la petite taille de son support spatial pour permettre une localisation précise des régions à décrire, et sa forme adaptée à la détection d'événements discriminants.

Covariance aux rotations. L'observation doit être indépendante du choix des axes du repère sur l'image I. Considérant que les coefficients transformés c_k sont paramétrés

sur $\mathbb{Z}^2 \times K$, la covariance de la transformée $T[I]$ de l'image $I \in \Omega_{nm}$ signifie :

$$\forall \theta \in [0, 2\pi[,\ \forall k_1 \in K,\ \exists k_2 \in K,\ T[I_\theta](\mathbf{x}, k_1) = T[I]\,(r_\theta(\mathbf{x}), k_2)\,,$$

où I_θ est l'image exprimée dans un repère tourné d'un angle θ, et r_θ la matrice de rotation usuelle. Les coefficients de l'image tournée n'appartiennent plus à une grille à coefficients entiers, et cela pose le problème de l'interpolation. Il est donc nécessaire de faire l'hypothèse que l'image et sa représentation sont échantillonnées à une fréquence supérieure à leur fréquence de Nyquist. Si l'ensemble K est réduit à un singleton, la covariance cherchée est équivalente à l'isotropie de h. Dans le cas contraire, la transformation $T[I]$ est définie sur $\mathbb{Z} \times [0, 2\pi[$ et est de la forme :

$$T[I](x, y; \theta) = [I * h_\theta](x, y) \tag{1.6}$$

où $h_\theta(x, y) = h(r_\theta(x, y))$. En pratique, la représentation n'est paramétrée que sur un ensemble fini d'orientations. Dans la section 4.5 seront introduites les transformations orientables, qui permettent à partir d'un nombre réduit d'orientations de générer par interpolation une représentation strictement covariante.

Covariance aux changements d'échelle. Il n'y a aucune information a priori sur l'échelle à laquelle apparaissent les objets qui composent la scène. Puisqu'on observe la scène à travers une représentation de la forme $R = I * h$, il est nécessaire d'adapter le support du filtre aux phénomènes observés. Les coefficients transformés sont donc paramétrés en échelle, conduisant à la transformation multi-échelles :

$$T[I](x, y; s) = [I * h_s](x, y) \tag{1.7}$$

où h_s est le filtre h dilaté par un facteur d'échelle s. La dilatation permet en effet de simuler une ouverture continûment croissante du filtre h (pour un filtre discret, cela pose le problème de l'interpolation de h sur une grille de résolution s fois plus grande). La condition nécessaire de représentation multi-échelle peut également s'obtenir en imposant à la représentation d'être indépendante du choix arbitraire de la résolution de l'image. Considérant une image échantillonnée à la fréquence de Nyquist, la condition de covariance de la transformée $T[I]$ aux changements de résolution s'exprime par :

$$\forall s > 1,\ \forall k_1 \in K,\ \exists k_2 \in K,\ T[I_s](\mathbf{x}, k_1) \ = \ T[I](\mathbf{x}, k_2) \tag{1.8}$$

où I_s est l'image I à une résolution s fois supérieure, soit d'après le théorème de Shannon, $I_s(\tilde{\mathbf{x}}) = [I * \mathrm{sinc}](\tilde{\mathbf{x}})$, où sinc est la fonction sinus cardinal séparable en x, y, et $\tilde{\mathbf{x}} = \frac{\mathbf{x}}{s}$ le point $\mathbf{x}$ de l'image I ramené sur la grille de l'image I_s. La relation 1.8 se réécrit :

$$\forall \mathbf{x} \in \mathbb{Z}^2,\ \forall s > 1,\ \forall s_1 > 1,\ \exists s_2 > s_1\ /\ [I * \mathrm{sinc}](\frac{\mathbf{x}}{s}) * h_{s_1}(\mathbf{x}) = [I * h_{s_2}](\mathbf{x}) \tag{1.9}$$

En particulier, cette relation est vérifiée pour $h_s(\mathbf{x}) = \mathrm{sinc}(\frac{\mathbf{x}}{s})$. En pratique, la condition de Nyquist n'est pas respectée et le choix des fonctions *sinc* comme fonctions d'interpolation n'est plus justifié. Il reste néanmoins nécessaire que le filtre h_{s_2} intègre la même information que h_{s_1} sur un support s fois plus grand. On retient donc comme modèle de filtre multi-échelle la famille de filtres générée par dilatation $\{h_s(\mathbf{x}) = h(\frac{\mathbf{x}}{s})\}_{s \in \mathbb{R}^{*+}}$.

1.1.3 Ondelettes et espace-échelle gaussien

Covariance des représentations en ondelettes. La contrainte de covariance aux changements arbitraires de repère a permis de restreindre les représentations possibles de l'image I aux convolutions de la forme :

$$T[I](\mathbf{y}, s, \theta) = \sum_{\mathbf{x} \in \mathbb{Z}^2} I(\mathbf{x}) h(s^{-1} r_\theta(\mathbf{y} - \mathbf{x})) \tag{1.10}$$

Dans le cas où le filtre h est isotrope, il n'y plus de dépendance en θ. La section 3.1.2 montrera que si le filtre h est une ondelette, c'est-à-dire s'il vérifie une condition d'admissibilité lui imposant d'être de moyenne nulle, la transformée continue par l'ondelette h d'une image I est très proche de la relation 1.10. La contrainte de covariance aux similitudes font des transformées continues en ondelettes des candidates naturelles pour de nombreux traitements visuels, comme la détection [Gro86] et la caractérisation [MH92] de contours, l'analyse d'images astronomiques [ADJ+02], l'analyse de turbulences 2D [Far92]. L'application de la transformée en ondelettes pour les problèmes de description d'images est plus rare. Les travaux dans ce domaine seront présentés dans la section 3.3. Ils ne concernent que la description globale et l'extraction de points d'intérêt. Il n'existe aucun schéma de description locale dans le domaine ondelettes. Cette absence de travaux dans ce domaine s'explique par l'hypothèse largement admise selon laquelle les représentations d'images adaptées à la description locale doivent être causales.

Représentations causales et espace-échelle gaussien. Pour des signaux monodimensionnels, une représentation multi-échelle est causale s'il n'y a pas pas création d'extrema locaux à échelle croissante. La causalité a été introduite dans [Wit83] pour modéliser une image de résolution décroissante, où les détails sont progressivement éliminés. Cette contrainte ne peut pas être satisfaite par les signaux multidimensionnels. L'extension naturelle proposée dans [Koe84] est d'imposer que les lignes de niveaux

$$\{(\mathbf{x}; s) \in \mathbb{R}^n \times \mathbb{R}^{*+} : T[I](\mathbf{x}; s) = cste\} \tag{1.11}$$

ne peuvent pas être créées à échelle croissante. La figure 1.3 montre un cas de création de maximum local respectant le principe de causalité. Les contraintes de linéarité, d'homogénéité, d'isotropie, et de causalité imposent à la représentation $R = T[I]$ de l'image I de vérifier l'équation de diffusion [Koe84] :

$$\frac{\partial R}{\partial s} = \frac{1}{2} \nabla R \tag{1.12}$$

dont l'unique solution est l'espace-échelle gaussien de l'image I.

Définition 5 *L'espace-échelle d'une image $I \in \Omega$ est la fonction L définie sur $\mathbb{Z}^2 \times \mathbb{R}^{*+}$ par :*

$$L(\mathbf{x}, s) = I * \frac{1}{2\pi s} e^{-\frac{||\mathbf{x}||^2}{2s}} \tag{1.13}$$

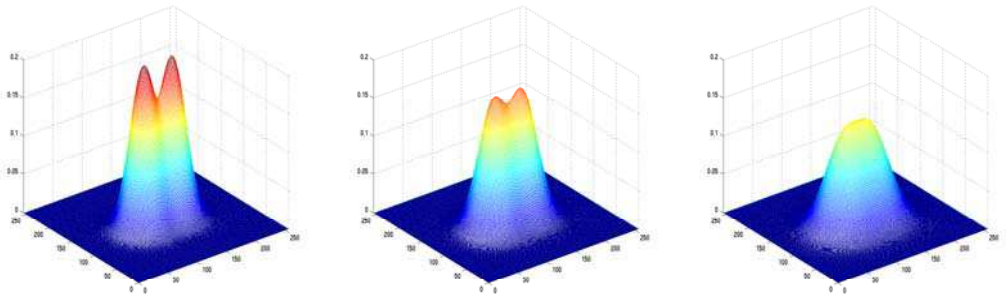

FIG. 1.3 – Exemple de création d'un maximum local dans une représentation causale (ici l'espace-échelle gaussien).

L'opération de dérivation sera fréquemment effectuée l'espace-échelle gaussien. La dérivée partielle d'une image I d'ordre $i+j$ sera notée $L_{x^i y^j}$ et définie par :

$$L_{x^i y^j} = \frac{\partial^{i+j} L}{\partial^i x \partial^j y} \tag{1.14}$$

Incertitude de localisation dans le plan espace-fréquence. L'information contenue par un coefficient d'une représentation obtenue par convolution linéaire porte sur une certaine région du plan espace-fréquence. Cette information est d'autant plus discriminante que cette région est petite. Il est donc intéressant dans un but de description, de chercher la représentation qui minimise la taille de cette région, relative à l'incertitude sur la localisation en espace et en fréquence de la réponse impulsionnelle. Une mesure simple de cette incertitude est le produit $\Delta_x.\Delta_\omega$ des écarts-type en espace et en fréquence de la réponse impulsionnelle. Ils sont définis par :

$$\begin{aligned} \Delta_x &= \sum_{\mathbf{x}\in\mathbb{Z}^2} ||\mathbf{x}-\mathbf{x_0}||^2 h_s(\mathbf{x}), \\ \Delta_\omega &= \sum_{\omega\in\mathbb{Z}^2} ||\omega-\omega_\mathbf{0}||^2 \hat{h}_s(\omega), \end{aligned} \tag{1.15}$$

$\mathbf{x_0}$ et $\omega_\mathbf{0}$ étant respectivement les moyennes en espace et en fréquence de la réponse impulsionnelle à l'échelle s, et $\hat{h}_s$ la transformée de Fourier h_s. Le noyau gaussien minimise la mesure d'incertitude définie par $\Delta_x.\Delta_\omega$. On verra dans la section 3.3.2 qu'il existe un lien étroit entre l'espace-échelle gaussien et les représentations par ondelettes continues, le laplacien de gaussienne étant l'ondelette faisant le pont entre les deux types de représentation.

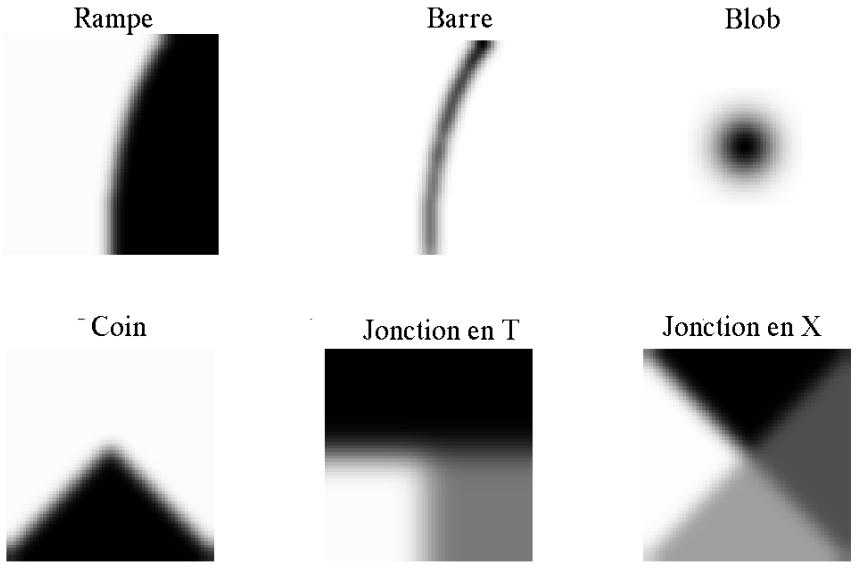

FIG. 1.4 – Exemples de primitives. Chacune d'entre elles possèdent des attributs caractéristiques qui peuvent être l'échelle, l'orientation, le contraste, la courbure.

1.2 Schémas de description locale

La section précédente a montré que pour s'affranchir du choix arbitraire du repère, les représentations (linéaires) d'images sont nécessairement multi-échelles, et isotropes ou multi-orientations. Le problème est désormais de savoir la nature de l'information à extraire et la forme qu'elle doit prendre. Le modèle le plus influent pour traiter ce problème est emprunté à la biologie décomposant la vision en étapes pré-attentive puis attentive [Nei64]. Dans l'étape pré-attentive sont uniquement captés les évènements visuels *saillants*. Ce sont des événements localisés, sur des *primitives saillantes* qui servent à initialiser le processus de vision à partir d'un nombre réduit d'échantillons. Les primitives les plus utilisées, répertoriées dans la figure 1.4, sont constituées d'une discontinuité spatiale de luminance. Le nombre d'attributs permettant de les caractériser varie selon le type de structure. Les attributs communs sont le contraste maximal, l'échelle précisant l'étendue de la variation dans la direction du gradient, la courbure de la surface de luminance dans cette direction. D'autres attributs sont spécifiques à certaines structures, comme l'orientation et la courbure d'une certaine ligne d'iso-luminance, l'échelle et le contraste d'une variation secondaire. Ces structures sont à haut pouvoir discriminant, donc rares, mais suffisamment génériques pour être partagées par toutes les images naturelles. Leur détection s'effectue durant la phase pré-attentive [HW62], et leur caractérisation durant la phase attentive. Cette deuxième phase consiste à regrouper les primitives extraites en fonction de leurs caractéristiques et de leurs relations spatiales particulières, puis à comparer les regroupements obtenus avec les prototypes connus. La difficulté de mise en oeuvre de ce modèle réside dans l'étape de regroupement ap-

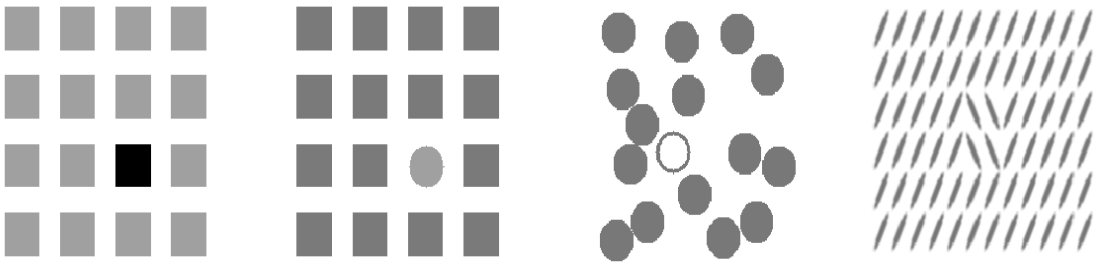

FIG. 1.5 – Exemples de saillances visuelles (adaptés de [KB01]).

paraissant dans la phase attentive. L'étape suivante de comparaison ne peut, en effet, fonctionner que si le regroupement reflète la structure de l'objet analysé, ce qui est très difficile sans connaissance a priori. Des travaux ont montré qu'il est en fait possible d'obtenir de bons taux de reconnaissance sans regroupement d'aucune sorte [SM97, SC00]. Les objets ne sont alors modélisés que par un ensemble de *descripteurs locaux* ne prenant pas en compte la position relative des régions décrites. Selon ce schéma, la description s'opère en deux étapes. La première étape est l'extraction des régions saillantes, et la seconde la description indépendante de chacune des régions extraites.

1.2.1 Extraction de régions saillantes

Deux conditions sur les régions extraites sont nécessaires pour que les descripteurs soient robustes et discriminants. Les régions extraites doivent, d'une part, être indépendantes des conditions de prise de vue et, d'autre part contenir suffisamment d'information. Pour extraire de telles régions, la quasi totalité des travaux utilisent le concept de saillance, et ne diffèrent que par le modèle de saillance utilisé. En vision biologique également, il en existe de nombreuses formes, comme le montre la figure 1.5. Les principaux modèles sont la saillance géométrique, la saillance entropique, et la saillance de symétrie.

Saillance géométrique. Le modèle de saillance géométrique regroupe les détecteurs de contours (rampes et barres), et les détecteurs de coins et de jonctions. Pour obtenir une description de haut niveau, les régions extraites doivent correspondre aux objets de la scène ou à leurs sous-parties homogènes. Ainsi, les premiers travaux en reconnaissance d'objets cherchaient à mettre en correspondance l'image binaire, obtenue par segmentation, avec les prototypes des objets modélisés. La difficulté à segmenter les longs contours, la fréquence élevée d'occultations dans les images naturelles, et la complexité algorithmique de la mise en correspondance, sont trois limites importantes de cette approche. Dans les travaux plus récents, l'extraction sans information a priori sur le contenu de la scène est beaucoup plus localisée, l'identification des objets ne se faisant qu'en fin du processus visuel [SM97]. Les régions extraites ne servent qu'à initialiser la

description, et sont des petits voisinages dont les centres constituent les points d'intérêt. Dans un modèle de saillance géométrique, les coins et les jonctions sont les primitives les plus utilisées. Contrairement aux contours, ils sont bien localisés, rendant leur extraction robuste à de nombreuses transformations. Ce sont également des événements rares, donc à haut pouvoir discriminant. Leur détection repose souvent sur les maxima locaux d'une mesure de saillance définie soit à partir du gradient et de la courbure, soit à partir d'une moyenne des variations des niveaux de gris dans toutes les directions. Dans tous les cas, la mesure de saillance est construite à partir de l'expansion au premier ou second ordre en série de Taylor de l'image I :

$$I(\mathbf{x}+\varepsilon) = I(\mathbf{x}) + \varepsilon^T \nabla I + \varepsilon^T H \varepsilon + O(\varepsilon^T \varepsilon) \tag{1.16}$$

où le gradient ∇I et la matrice Hessienne H sont définis à partir des dérivées partielles de l'équation 1.14 par :

$$\nabla I = \begin{pmatrix} L_x \\ L_y \end{pmatrix}, \quad H = \begin{pmatrix} L_{xx} & L_{xy} \\ L_{yx} & L_{yy} \end{pmatrix} \tag{1.17}$$

Dans leur formulation originelle, certains des détecteurs présentés dans cette section estimaient les dérivées partielles par simple différence entre pixels voisins de l'image. L'opération de dérivation est ici choisie dans l'espace-gaussien. Cela introduit un nouveau paramètre, l'échelle de dérivation. Les détecteurs peuvent alors être mono-échelle si la détection se fait à une échelle fixée, ou multi-échelles dans le cas contraire. Le problème de la sélection d'une échelle robuste pour une extraction multi-échelles est traité dans la section 2.1.2.

Beaudet définit dans [Bea78] la mesure de saillance E des coins et des jonctions comme le déterminant de la matrice Hessienne, qui est lié au produit $k_{min}k_{max}$ des courbures principales et de la norme du gradient par [Lip69] :

$$E = Det(H) = k_{min}k_{max}(1 + {L_x}^2 + {L_y}^2)^2$$

Deriche et Faugeras calculent cette mesure à deux échelles différentes, et un gain en localisation est obtenu en détectant le passage à zéro du laplacien le long de la droite définie par les deux coins extraits [DF90]. Une mesure similaire, proposée Kitchen et Rosenfeld dans [KR82], est le produit de la norme du gradient et de la dérivée de l'orientation du gradient le long du contour. Elle s'écrit simplement à partir des dérivées partielles :

$$E = \frac{L_{xx}{L_y}^2 - 2L_{xy}L_xL_y + L_{yy}{L_x}^2}{{L_x}^2 + {L_y}^2}$$

Dans [Mor77], Moravec introduit les « points d'intérêt », définis par les maxima locaux d'une mesure de la variation bidimensionnelle des niveaux de gris. En tout point (x, y) de l'image est définie l'énergie orientée selon le vecteur de déplacement $\mathbf{d} = (d_1, d_2)$:

$$E_{x,y}(\mathbf{d}) = \sum_{u_1,u_2} w(u_1,u_2)\ |I(x+u_1+d_1, y+u_2+d_2)\ -\ I(x+u_1, x+u_2)|^2$$

où w est une fenêtre rectangulaire. Le calcul de l'énergie orientée pour un nombre suffisant de déplacements permet d'identifier la structure locale (x, y). Harris propose dans [HS88] une méthode pour accéder à moindre coût aux directions d'intérêt, que sont celles conduisant aux variations minimales et maximales. Introduisant une fenêtre gaussienne w_s (d'écart-type s) et l'expansion en série de Taylor de l'image I, l'énergie orientée se réécrit :

$$E_{x,y,s}(\mathbf{d}) = \sum_{u_1,u_2} w_s(u_1,u_2) \left| d_1 \frac{\partial I}{\partial x}(x+u_1, y+u_2) \; + \; d_2 \frac{\partial I}{\partial y}(x+u_1, y+u_2) \; + \; O(d_1^2, d_2^2) \right|^2$$

où le nouveau paramètre s est une échelle d'intégration. La symétrie de la gaussienne permet d'écrire cette énergie sous la forme d'une convolution :

$$E_{x,y,s}(\mathbf{d}) = \mathbf{d}\, M(\mathbf{x}, s)\, \mathbf{d}^T \tag{1.18}$$

où la matrice $M(\mathbf{x}, s)$ est la convolution entre ω_s et la matrice d'autocorrélation des dérivées partielles de premier ordre :

$$M(\mathbf{x}, s) = \begin{pmatrix} w_s * {L_x}^2 & w_s * L_x L_y \\ w_s * L_x L_y & w_s * {L_y}^2 \end{pmatrix} \tag{1.19}$$

Les déplacements causant les variations minimales et maximales sont donc les vecteurs propres de la matrice d'autocorrélation des dérivées partielles de premier ordre. C'est pourquoi la matrice $M(\mathbf{x}, s)$ est aussi appelée matrice de structure en raison de l'information qu'elle porte sur la structure locale en $\mathbf{x}$. Deux valeurs propres élevées signifient qu'il existe au moins deux directions de forte variation, et donc que le point est proche d'un coin ou d'une jonction. Une seule valeur propre élevée signifie que le point est proche d'un contour. La détection simultanée de coins et de contours peut donc s'effectuer par la recherche des maxima locaux de l'énergie de Harris définie par :

$$E_\alpha(\mathbf{x}, s) = |det(M(\mathbf{x}, s)) \; - \; \alpha\, trace^2(M(\mathbf{x}, s))| \tag{1.20}$$

La détection des seuls coins et jonctions peut s'effectuer par la recherche des maxima locaux de la mesure de Förstner, définie comme la plus petite valeur propre de la matrice de structure.

Saillance entropique. Dans un modèle de saillance entropique, les points extraits sont les points dont le voisinage contient un maximum d'information. La description des régions ainsi extraites dispose d'un fort potentiel discriminant. La mesure de saillance, proposée par Khadir et Brady dans [KB01], pour extraire de tels points est définie par le produit entre l'entropie locale et une mesure d'auto-dissimilarité :

$$E(\mathbf{x}, s) = X(\mathbf{x}, s) Y(\mathbf{x}, s)$$

L'entropie locale $X(\mathbf{x}, s)$ est estimée par :

$$X(\mathbf{x}, s) = \sum_i p_{\mathbf{x},s}(i) \log\left(p_{\mathbf{x},s}(i)\right)$$

où $p_{\mathbf{x},s}(i)$ est la fréquence du niveau de gris i dans le voisinage centré en $\mathbf{x}$ et de taille s. Les régions bruitées ont une entropie locale élevée et peuvent s'éliminer en constatant que la distribution du bruit est auto-similaire à travers les échelles. La mesure d'auto-dissimilarité $Y(\mathbf{x}, s)$ est définie par :

$$Y(\mathbf{x}, s) = s \sum_i |\frac{\partial}{\partial s} p_{\mathbf{x},s}(i)|$$

Il existe en fait un lien étroit entre saillance géométrique et saillance entropique : les points détectés par saillance entropique sont situés en des régions à forte variation bidimensionnelle.

Saillance de symétrie. À partir de considérations psycho-visuelles, il est possible de chercher à rendre saillantes les structures symétriques. Une structure intéressante pour la description est le « blob »de luminance modélisé par une gaussienne. Une image présentant un blob en (x_0, y_0) peut être localement modélisée par :

$$I(x_0 + u_1, y_0 + u_2) = a + b\, e^{-\frac{u_1^2 + u_2^2}{2t^2}} \tag{1.21}$$

dont l'écart-type t est l'échelle caractéristique. Le laplacien est maximal au centre du blob. Lindeberg propose dans [Lin94b] d'extraire les blobs en détectant les maxima locaux (en espace et en échelle) du laplacien normalisé présenté dans la section 2.1.2.

D'autres mesures de saillance symétrique existent. Dans [Res95], la saillance est définie en coordonnées polaires par :

$$S(\mathbf{r_0}) = \sum_{r,\alpha} |\nabla I(r_0 + r, \alpha)||\nabla I(r_0 - r, \alpha)| D_r P_0(r, \alpha)$$

où les gradients pris en deux points radialement symétriques par rapport à $\mathbf{r_0}$ sont pondérés par la gaussienne D_r centrée en $\mathbf{r_0}$ et de variance $2||\mathbf{r}||$, et par P_0 la mesure de symétrie proprement dite au point $\mathbf{r_0}$. Elle est définie par :

$$P_0(r, \alpha) = [1 - \cos\left(\theta_1 + \theta_2 - 2\alpha\right)][1 - \cos\left(\theta_1 - \theta_2\right)]$$

où θ_1, θ_2 sont les orientations du gradient aux points radialement symétriques par rapport à $\mathbf{r_0}$. Dans [Kov97], la saillance est une mesure de la congruence des harmoniques de Fourier. Il est, en effet, constaté que les axes de symétrie apparaissent lorsque toutes les harmoniques sont en phase en un extremum, et les axes de dissymétrie lorsque les harmoniques s'annulent simultanément. Bigun a montré que la matrice de structure définie en 1.19 peut permettre la détection de nombreuses primitives symétriques si l'on utilise des opérateurs de dérivation complexes [Big04].

1.2.2 Description des régions saillantes

Étant donnée la famille $\{(\mathbf{x_i}, s_i, \theta_i)\}_{1 \leq i \leq n}$ des caractéristiques (points, échelles, et orientations) extraites à partir de la représentation $R(\mathbf{x}, s, \theta)$ d'une image I, le problème consiste désormais à décrire chacun des voisinages V_i des points d'intérêt. Dans la section précédente, la représentation R est isotrope, et la famille des caractéristiques extraites est réduite aux points d'intérêt $\{\mathbf{x_i}\}_{1 \leq i \leq n}$, ou aux points et à leurs échelles caractéristiques $\{(\mathbf{x_i}, s_i)\}_{1 \leq i \leq n}$. Selon le type d'extraction, mono ou multi-échelle, la description a lieu dans l'espace image, ou dans l'espace-échelle gaussien. Dans les prochains chapitres, l'extraction et la description pourront avoir lieu dans des espace-échelles orientés.

L'espace des descripteurs est un espace vectoriel normé, où une distance permet de mesurer la similarité entre les descripteurs. Un descripteur doit être invariant aux conditions de pose, et porter la spécificité de la structure présente dans le voisinage extrait. Les deux contraintes sont antagonistes : l'invariance requise pour la robustesse impose aux descripteurs d'appartenir à un sous-ensemble, ce qui a pour effet de réduire leur pouvoir discriminant. Formellement, le problème de détection de copies n'est pas celui de la reconnaissance d'images d'une même scène. Au contraire, il souhaitable de ne pas détecter comme copies deux images différentes d'une même scène. En pratique néanmoins, l'élimination de ces fausses alarmes nécessite de détecter et de comparer l'axe optique de chacune des images, ce qui engendre un coût important en temps de calcul.

Le descripteur le plus simple est la matrice des valeurs de luminance dans le voisinage extrait. La mesure de similarité est alors une corrélation. Ce descripteur est de grande dimension, ce qui restreint les applications possibles, et n'est pas invariant aux rotations. Cette section présente trois types de descripteur : les descripteurs calculés à partir d'invariants scalaires dans l'espace image ou dans l'espace-échelle ; les descripteurs d'*apparence* consistant en la répartition de caractéristiques ou d'attributs préalablement estimés ; les descripteurs d'*apparence* et de *forme*.

Descripteurs calculés à partir de scalaires invariants. Les invariants peuvent se calculer directement dans la représentation en niveaux de gris, comme les invariants algébriques, ou dans des représentations permettant la construction d'invariants plus discriminants. Les invariants algébriques sont des combinaisons de moments invariantes aux similitudes. Le moment $m^i_{p,q}$ d'ordre $p+q$ du voisinage V_i centré en (x_i, y_i), défini par :

$$m^i_{p,q} = \int_{x,y \in V_i} (x - x_i)^p (y - y_i)^q I(x, y) dx dy$$

est invariant aux translations (si les points extraits le sont). L'invariance aux changements d'échelle s'obtient en introduisant les moments normalisés :

$$\mu^i_{p,q} = \frac{m^i_{p,q}}{{m^i_{0,0}}^{1+(p+q)/2}}$$

L'invariance aux rotations s'obtient en formant les moments de Hu [Hu62] par combinaisons linéaires adéquates de moments normalisés. Les moments $m^i_{p,q}$ sont les projections de l'image I sur les polynômes $(x - x_i)^p (y - y_i)^q$. Les moments de Zernike sont les projections sur des polynômes orthogonaux, conduisant à des moments décorrélés dont le pouvoir discriminant est plus grand.

Des invariants peuvent être calculés dans l'espace-fréquence d'une image I, donné par une transformée comme celle de Fourier-Mellin définie en coordonnées polaires, pour $\sigma > 0$ fixé, par :

$$\forall (k,u) \in \mathbb{Z} \times \mathbb{R},\ T_\sigma[I](k,u) = \frac{1}{2\pi} \int_0^\infty \int_0^{2\pi} I(r,\theta) r^{\sigma - iu} e^{-ik\theta} \frac{dr}{r} d\theta$$

Cette transformée est bien adaptée au calcul d'invariants, chacune des harmoniques de Fourier-Mellin normalisées

$$D_\sigma[I](k,u) = T_\sigma(0,0)^{\frac{-\sigma+iu}{\sigma}} e^{ik \arg(M_\sigma(1,0))} T_\sigma(k,u)$$

étant invariante aux similitudes [Gho94]. Ces harmoniques normalisées peuvent s'appliquer à la description locale : la région V_i associée au point $(\mathbf{x_i}, s_i)$ est décrite par l'ensemble $\{D_{s_i}[I_{V_i}](k,u)\}_{(k,u)\in\mathbb{Z}\times\mathbb{R}}$, où I_{V_i} est la restriction de l'image I sur le voisinage V_i.

Les invariants différentiels, introduits dans [KD87], ont été les premiers appliqués à la description locale [SM96]. Contrairement aux précédents invariants, ils ne sont invariants qu'au groupe des translations et des rotations. Dans [SM97], une représentation multi-échelle est utilisée pour calculer les invariants à l'échelle caractéristique des points extraits, et ainsi obtenir l'invariance aux changements d'échelle. Les invariants différentiels sont des combinaisons de polynômes homogènes et symétriques de dérivées partielles, comme par exemple :

$$\begin{bmatrix} L \\ L_i L_i \\ L_i L_{ij} L_j \\ L_{ii} \\ L_{ij} L_{ji} \\ \varepsilon_{ij}(L_{jkl} L_i L_k L_l - L_{jkk} L_i L_l L_l) \\ L_{iij} L_j L_k L_k - L_{ijk} L_i L_j L_k \\ -\varepsilon_{ij} L_{jkl} L_i L_k L_l \\ L_{ijk} L_i L_j L_k \end{bmatrix} \tag{1.22}$$

où la notation de Einstein est utilisée ($L_{ii} = \sum_{i\in\{x,y\}} L_{ii} = L_{xx} + L_{yy}$ par exemple), et où le tenseur ε est défini par $\varepsilon_{xx} = \varepsilon_{yy} = 0$ et $\varepsilon_{xy} = -\varepsilon_{yx} = 1$. Ils sont calculés en chacun des points extraits, à leur échelle caractéristique telle que définie dans 2.1.2. Chaque invariant intègre les variations locales du signal image autour d'un point extrait $(\mathbf{x_i}, s_i)$, et porte donc une information sur la géométrie du voisinage V_i, ce qui lui confère son pouvoir discriminant.

Descripteurs calculés à partir d'une distribution invariante. La deuxième famille de descripteurs est composée des distributions d'un attribut caractéristique (comme l'orientation, ou la courbure) dans le voisinage de chaque point extrait. Une attention particulière doit être portée sur la pondération des éléments contribuant à l'histogramme, et à la normalisation de l'histogramme. L'invariance aux changements d'échelle s'obtient aisément en adaptant à l'échelle caractéristique la taille du voisinage sur lequel est calculé l'histogramme. Ce type de description remonte aux histogrammes de couleur proposés dans [SB91]. D'autres attributs que la couleur, comme l'intensité des réponses obtenues par des dérivées de gaussienne ou par des filtres de Gabor, sont proposés dans [SC96]. Ces histogrammes 1D ne prennent pas en compte la répartition spatiale des attributs dans le voisinage du point, ce qui leur confère l'invariance aux rotations au prix d'une limitation de leur pouvoir discriminant. Les histogrammes 2D ou 3D donnant la répartition spatiale d'un attribut sont plus discriminants. Dans ce cas, la robustesse à la rotation peut s'obtenir de deux manières différentes. La plus simple consiste à ne considérer que la répartition radiale, conduisant à des histogrammes 2D dont les axes sont la valeur de l'attribut considéré et la distance au point d'intérêt. Lorsque l'attribut considéré est la luminance, ce descripteur, appelé « spin image », est proposé dans [JH99] pour la mise en correspondance de surfaces 3D. La robustesse aux variations monotones de luminance est obtenue en modifiant la matrice ainsi construite pour être de moyenne nulle et de norme de Froebenius unité. Dans [YF02], la même méthode est utilisée avec la courbure comme attribut. La robustesse à la rotation peut également s'obtenir en détectant une orientation robuste en tout point extrait, permettant de construire des histogrammes 3D encore plus discriminants. Les axes sont alors la valeur de l'attribut et les deux directions perpendiculaires dont l'une est fixée par l'orientation du point central. Le descripteur SIFT [Low99] construit de cette manière avec l'orientation du gradient comme attribut, est à ce jour le descripteur le plus robuste et le plus discriminant [MS05]. Il sert de référence et est analysé en détail dans la section 2.4.

Descripteurs d'apparence et de forme. Pour la reconnaissance d'objets, les descripteurs doivent discriminer différentes classes d'objets tout en tolérant une variabilité éventuellement forte des objets dans leur classe. Dans ce contexte, la position relative des contours est une information importante permettant à la fois d'augmenter le pouvoir discriminant et de modéliser la variabilité inter-classe des objets. L'information à prendre en compte dans le modèle des objets ne doit pas se réduire comme dans les applications précédentes à *l'apparence*, mais doit également porter sur *la forme*. Les premiers modèles utilisés en reconnaissance ne portaient que sur la forme, décrite à partir de l'image segmentée. L'instabilité de la segmentation et la difficulté à mettre en correspondance des images binaires sont deux obstacles majeurs aux approches uniquement basées sur la forme. Les efforts actuels portent sur l'élaboration de modèles incorporant simultanément la forme et l'apparence. Dans [BMP02], le descripteur « spin image »est modifié pour prendre en compte une information de forme en ne faisant contribuer à

l'histogramme final que les pixels du voisinage détectés sur un contour. Dans [MHS05], le descripteur SIFT est enrichi d'un histogramme donnant la répartition polaire dans un voisinage très large des pixels situés sur un contour. Enfin, dans [WWP00] a été proposé un schéma de reconnaissance d'objets reposant sur un modèle probabiliste qui a influé sur de nombreux travaux. Dans ce schéma, un modèle est un assemblage de parties rigides, dotées de positions relatives variables. La position de chaque élement est representée par une densite de probabilites, calculée dans la phase d'apprentissage.

1.3 Évaluation de la description locale

Il n'existe pas de protocole partagé par tous les scientifiques travaillant dans ce domaine. L'évaluation de la performance du schéma complet de description s'effectue généralement par requêtes par le contenu. La vérité terrain entre les requêtes et la base d'images est connue : les requêtes sont soit des copies synthétiques d'images de la base, soit des images d'une même scène. La base d'images est selon l'application, soit une base d'images hétérogènes, soit une base d'images homogènes comprenant des empreintes, des visages, des objets sur fond uni et fixe, ou sur fond texturé et variable. L'évaluation de la première étape du schéma de description, l'extraction de points et éventuellement d'échelles et d'orientations, est plus objective. La première section présente les techniques d'évaluation de la robustesse des caractéristiques extraites, la seconde les techniques d'évaluation de la performance des descripteurs locaux.

1.3.1 Évaluation de l'extraction

Le concept de saillance a été introduit pour extraire des régions robustes et discriminantes. Le pouvoir discriminant peut être évalué indépendamment de la description, au moyen de l'entropie locale comme dans [SMB98], ou conjointement avec la description, au moyen de la méthode décrite dans la section 1.3.2. Le but de cette section est de définir un protocole expérimental permettant d'évaluer la robustesse des points extraits.

Restriction des transformations admissibles pour l'évaluation. Les régions extraites sont des voisinages de petite taille centrés sur les points d'intérêt. La scène correspondant à ces régions est donc approximativement plane. La description visant à obtenir de l'information intrinsèque à la scène, donc indépendante des conditions de prise de vue, les régions extraites sont idéalement robustes à toute transformation photométrique et homographique. Approchant les transformations homographiques par des transformations affines, les régions extraites doivent être robustes aux transformations monotones en luminance et affines en spatial. Hormis les problèmes de saturation, les transformations photométriques monotones ne modifient pas la structure locale des images et la robustesse à ce type de transformation s'obtient aisément. La robustesse aux transformations spatiales affines est plus délicate. L'évaluation effectuée dans [MS05] montre que les extracteurs conçus pour être robustes aux transformations affines ne sont, en

fait, pas meilleurs que les extracteurs robustes aux seules similitudes. Les *transformations admissibles* pour l'évaluation de la description sont restreintes aux similitudes, aux crops, et aux compressions JPEG. En effet, l'expérience montre qu'une description robuste à ces quelques transformations admissibles est robuste à un ensemble beaucoup plus vaste de transformations. La transformation admissible utilisée pour l'évaluation commune de tous les extracteurs présentés dans cette thèse est composée d'une dilatation de facteur 1.6, d'une rotation d'angle 30 degrés, d'un crop de facteur 30%, et d'une compression JPEG de facteur 10 (0.8 bit/pixel). C'est à cette transformation que fera référence une copie non définie autrement par le contexte. Précisons enfin que le choix des fonctions d'interpolation pour créer la copie n'a que peu d'influence sur l'évaluation de la robustesse. La figure 1.6 montre l'image Lena et sa copie.

Sélection des points d'intérêt pour l'évaluation. D'après la figure 1.6 l'évaluation de la robustesse des points extraits doit mesurer la correspondance, pour une transformation admissible donnée, entre les points $(\mathbf{x_i}, s_i, \theta_i)$, extraits à partir d'une image I, et les points $(\breve{\mathbf{x}}_\mathbf{j}, \breve{s}_j, \breve{\theta}_j)$, extraits à partir d'une image transformée I_t. La transformée synthétique entre I et I_t étant connue, les coordonnées $(\breve{\mathbf{x}}_\mathbf{j}, \breve{s}_j, \breve{\theta}_j)$ sont considérées non pas dans le repère de I_t mais ramenées dans celui de I (c'est-à-dire que $\breve{s}_j$ et $\breve{\theta}_j$ prennent respectivement en compte le changement de résolution et l'angle de rotation entre l'image originale et la copie). Certaines transformations admissibles, comme les crops, font que le support de la copie ne coïncide pas avec celui de l'image originale. L'évaluation de la correspondance se fait donc uniquement à partir des points appartenant à l'intersection de ces deux supports, notée S. Dans la suite, les familles $\{\mathbf{x_i}\}_{1\leq n_1}$ et $\{\breve{\mathbf{x}}_\mathbf{j}{}_{1\leq n_2}$ sont constituées des points appartenant à ce support commun.

Critères d'évaluation de la robustesse des caractéristiques extraites. La robustesse des caractéristiques extraites peut être évaluée chacune indépendamment ou plusieurs simultanément.

Définition 6 *La répétabilité surfacique r_s est le ratio entre la surface d'intersection des régions extraites et la surface de leur union [SMB98] :*

$$r_s = \frac{|V \cap V_t|}{|V \cup V_t|} \tag{1.23}$$

où $|V|$ est le cardinal de l'ensemble V, et où V et V_t sont l'union des régions extraites à partir de I et I_t, et incluses dans le support S commun aux deux images.

La région extraite autour d'un point $\mathbf{x_i}$ étant un carré ou un cercle de taille liée à l'échelle s_i du point, ce critère évalue simultanément la robustesse des points et la robustesse des échelles. Toutefois, ce critère d'évaluation dépend fortement de la taille des régions extraites. La répétabilité surfacique atteint rapidement la répétabilité maximale lorsque les régions extraites croissent.

FIG. 1.6 – Visualisation d'une copie composée d'une dilatation de facteur 1.6, d'une rotation d'angle 30 degrés, d'un crop de facteur 30%, et d'une compression JPEG de facteur 10. Les carrés ont leur centre en les points extraits par le détecteur de Harris-Laplace, et leur coté de taille proportionnelle à l'échelle du point extrait.

Dans cette thèse, la robustesse des caractéristiques s'effectuera par trois mesures : la répétabilité r_p des points seuls, la répétabilité $r_{p,s}$ des points et des échelles, et la répétabilité $r_{p,s,\theta}$ des points, des échelles et des orientations. Chacune des répétabilités sera évaluée à une précision donnée.

Définition 7 *Soit* $(\mathbf{x}_i, s_i, \theta_i)$ *et* $(\breve{\mathbf{x}}_j, \breve{s}_j, \breve{\theta}_j)$ *deux triplets de points, échelles, orientations, extraits respectivement sur l'image originale et l'image transformée (ramenés dans le repère original, et appartenant au support commun).*
L'erreur de localisation est

$$\varepsilon = ||\mathbf{x}_i - \breve{\mathbf{x}}_j||_2 \tag{1.24}$$

L'erreur d'estimation en échelle est

$$\varepsilon_s = \frac{\max(s_i, \breve{s}_j)}{\min(s_i, \breve{s}_j)} \tag{1.25}$$

L'erreur d'estimation en orientation est

$$\varepsilon_\theta = |\theta_i - \breve{\theta}_j| \tag{1.26}$$

Ainsi si $\varepsilon = 0$ et si $\varepsilon_s < 2$, les points $(\mathbf{x}_i, s_i)$ et $(\breve{\mathbf{x}}_j, \breve{s}_j)$ coïncident et appartiennent à la même octave.

Définition 8 *La répétabilité* $r_p(\varepsilon)$ *est la proportion de paires de points qui se correspondent à une erreur* ε *près relativement au nombre maximum de paires qui peuvent se correspondre [SMB98] :*

$$r_p(\varepsilon) = \frac{|\{i \in [\![1, n_1]\!] \,|\, \exists j \in [\![1, n_2]\!] \,:\, ||\mathbf{x}_i - \breve{\mathbf{x}}_j||_2 \leq \varepsilon\}|}{\min(n_1, n_2)} \tag{1.27}$$

Pour prendre en compte la robustesse des échelles extraites, la répétabilité des points est adaptée de la manière suivante :

Définition 9 *La répétabilité* $r_{p,s}(\varepsilon, \varepsilon_s)$ *est la proportion de paires de points qui se correspondent à une erreur de localisation* ε *et une erreur d'estimation en échelle* ε_s *près :*

$$r_{p,s}(\varepsilon, \varepsilon_s) = \frac{|\{i \in [\![1, n_1]\!] \,|\, \exists j \in [\![1, n_2]\!] \,:\, ||\mathbf{x}_i - \breve{\mathbf{x}}_j||_2 \leq \varepsilon, \exp(|\log(\frac{s_i}{\breve{s}_j})|) \leq \varepsilon_s\}|}{\min(n_1, n_2)} \tag{1.28}$$

De même que la répétabilité $r_p(\varepsilon)$ a été adaptée en $r_{p,s}(\varepsilon, \varepsilon_s)$ pour prendre en compte la robustesse des échelles extraites, la répétabilité $r_{p,s,\theta}(\varepsilon, \varepsilon_s, \varepsilon_\theta)$ évalue le taux de points correctement détectés à une erreur de localisation ε près, et des erreurs d'estimation en échelle ε_s et en orientation ε_θ près.

Influence du type d'image et du nombre de points sur les répétabilités. Les évaluations faites dans [SMB98, MS05] ne mentionnent pas le nombre de points extraits ni la taille des images utilisées. Les différentes répétabilités définies ci-dessus sont pourtant très sensibles à ces deux facteurs. L'espérance de la répétabilité définie par la relation 1.27 d'un tirage aléatoire de n points à partir d'une image composée de N pixels vaut :

$$r_p(\varepsilon) = \frac{n\varepsilon^2}{N}$$

Dans cette thèse, le nombre de points utilisés pour évaluer les répétabilités est compris entre 100 et 200 points selon les images ; et les images sont de taille comprise entre 512×380 pixels et 512×512 pixels. Enfin, les répétabilités mesurées sur différentes images peuvent varier d'un facteur de 1 à 4 selon les images. Les résultats sont donc moyennés à partir de 100 images tirées aléatoirement dans une grande base d'images hétérogènes.

1.3.2 Évaluation de la description

Travaux existants. Les travaux sur l'évaluation de la description sont très peu nombreux. Aucune méthode n'a été proposée pour évaluer les descripteurs indépendamment de la robustesse des points extraits. La seule méthode d'évaluation existante consiste à mesurer la capacité des descripteurs à apparier correctement les points d'images d'une même scène. L'évaluation se fait pour une base d'images fixée, à partir d'images requêtes créées par copies d'images de la base. Les points extraits sur la copie sont appariés avec les points de la base dont les descripteurs sont distants de moins d'un rayon r que le descripteur requête. Le rappel est le nombre d'appariements corrects divisé par le nombre total d'appariements :

$$\textit{rappel} = \frac{\text{nombre d'appariements corrects}}{\text{nombre d'appariements}} \tag{1.29}$$

où un appariement est jugé correct si la distance entre les points appariés est inférieure à un seuil. Ce seuil est fixé dans toute l'évaluation, et est en général égal à $\varepsilon = 1.5$ pixels. L'évaluation consiste en la courbe de précision-rappel construite en faisant varier le rayon r définissant l'ensemble des appariements pour un descripteur requête fixé. La précision est le nombre de points correctement appariés divisé par le nombre total de points extraits :

$$\textit{précision} = \frac{\text{nombre d'appariements corrects}}{\text{nombre de points extraits}} \tag{1.30}$$

Ainsi, pour des grands rayons r d'appariement, le rappel est petit, et la précision grande. Pour chaque point extrait sur la copie, le descripteur le plus proche peut être cherché soit uniquement dans l'ensemble des descripteurs de l'image originale [KS04], soit dans l'ensemble des descripteurs d'une base d'images [CJ02].

Méthode proposée. Dans cette thèse, l'évaluation des descripteurs est effectuée par une nouvelle méthode permettant de s'affranchir de la robustesse des points extraits. L'évaluation est effectuée par le taux T_d d'appariements corrects par descripteur. Se donnant le support commun S entre l'image originale et l'image transformée, et considérant l'ensemble des points $\{\mathbf{x_i} \in S\}_{1 \leq i \leq n_0}$ extraits sur l'image originale I et l'ensemble des points $\{\breve{\mathbf{x}}_\mathbf{j} \in S\}_{1 \leq j \leq n_t}$ (ramenés dans le repère de I) extraits à partir de la copie I_t, une méthode d'évaluation des descripteurs consiste, contrairement au cas précédent, à n'apparier les points de l'image transformée qu'avec ceux de l'image originale dont le descripteur est le plus proche.

Définition 10 *La répétabilité par descripteurs est définie par :*

$$r_d(\varepsilon) = \frac{|\{\breve{\mathbf{x}}_\mathbf{j} \in S : ||\breve{\mathbf{x}}_\mathbf{j} - \mathbf{x_{c(j)}}||_2 \leq \varepsilon\}|}{\min\left(|\{\mathbf{x_i} \in S\}|, |\{\breve{\mathbf{x}}_\mathbf{j} \in S\}|\right)} \tag{1.31}$$

où c est la fonction d'appariement au sens du descripteur le plus proche et donc définie par :

$$\forall 1 \leq j \leq n_t,\ c(j) = \arg\min_{1 \leq i \leq n_0} ||\breve{\mathbf{d}}_\mathbf{j} - \mathbf{d_i}||_2$$

$\{\mathbf{d_i}\}_{1 \leq i \leq n_0}$ *et* $\{\breve{\mathbf{d}}_\mathbf{j}\}_{1 \leq j \leq n_t}$ *étant respectivement les descripteurs calculés en* $\{\mathbf{x_i}\}_{1 \leq i \leq n_0}$ *et en* $\{\breve{\mathbf{x}}_\mathbf{j}\}_{1 \leq j \leq n_t}$

Contrairement aux travaux existants, un seul point de la copie est apparié avec un point donné de l'image originale.

Définition 11 *Soit $\breve{\mathbf{x}}_\mathbf{j}$ un point d'intérêt extrait sur une copie. Ce point est apparié au sens du descripteur le plus proche avec le point d'intérêt $\mathbf{x_{c(j)}}$ extrait sur l'image originale. L'appariement est dit correct à ε près lorsque la distance $||\breve{\mathbf{x}}_\mathbf{j} - \mathbf{x_{c(j)}}||_2$ entre ces deux points est inférieure à ε pixels.*

À ε fixé, la répétabilité par descripteurs $r_d(\varepsilon)$ est majorée par la répétabilité $r_p(\varepsilon)$ définie en 1.27.

Définition 12 *Le taux d'appariements corrects à ε près est le ratio :*

$$T_d(\varepsilon) = \frac{r_d(\varepsilon)}{r_p(\varepsilon)} \tag{1.32}$$

Il s'agit de la probabilité que le point apparié par le descripteur le plus proche soit distant de moins de ε pixels, sachant qu'un tel point existe. Ce taux permet d'évaluer la performance de la description indépendamment de la robustesse des points extraits.

Mesure de la sensibilité des descripteurs aux erreurs de détection. Il est intéressant de connaître les causes d'un mauvais appariement par descripteur le plus proche. Le descripteur utilisé dans cette thèse est le descripteur SIFT [Low99] présenté dans la

section 2.4. Pour être calculé, il requiert l'extraction préalable de points, d'échelles et d'orientations. Un mauvais appariement peut être dû à une erreur de localisation du point d'intérêt, ou à une erreur d'estimation en échelle ou en orientation. Il est important de mesurer la sensibilité du descripteur à ces erreurs d'estimation. La mesure de cette sensibilité permet de connaître l'erreur maximale qu'il est possible de s'autoriser durant la phase d'extraction des caractéristiques.

Soit $j \in [\![1, n_t]\!]$ fixé. Le point $\check{\mathbf{x}}_\mathbf{j}$ de l'image transformée est apparié avec le point $\mathbf{x}_i$ de l'image originale par le descripteur le plus proche. Par commodité, les notations suivantes seront fréquemment utilisées dans la thèse :

- La notation $P(\text{bon appariement}|\varepsilon)$ réfère à la probabilité d'un appariement correct à ε près, sachant qu'il existe un point de l'image originale dont l'erreur de localisation est inférieure à ε, et dont les erreurs d'estimation en échelle et en orientation sont inférieures à 1.3 et 15 degrés.
- La notation $P(\text{bon appariement}|\varepsilon_s)$ réfère à la probabilité d'un appariement correct à 1.5 pixels près, sachant qu'il existe un point de l'image originale dont l'erreur de localisation est inférieure à 1.5 pixels, et dont les erreurs d'estimation en échelle et en orientation sont inférieures à ε_s et 15 degrés.
- La notation $P(\text{bon appariement}|\varepsilon_\theta)$ réfère à la probabilité d'un appariement correct à 1.5 pixels près, sachant qu'il existe un point de l'image originale dont l'erreur de localisation est inférieure à 1.5 pixels, et dont les erreurs d'estimation en échelle et en orientation sont inférieures à 1.3 et ε_θ degrés.

Ces quantités permettent de mesurer la sensibilité des descripteurs à une erreur donnée, toute autre erreur étant négligeable par ailleurs. Le choix de l'erreur maximale jusqu'à laquelle une erreur peut être considérée négligeable est justifié dans le prochain chapitre.

1.4 Conclusion

Ce chapitre a dans un premier temps restreint considérablement les représentations (linéaires) possibles pour le seul problème de description. En vue de s'affranchir du choix arbitraire du repère sur lequel sont définies les images, les représentations doivent être covariantes aux similitudes, donc nécessairement multi-échelles, et isotropes ou multi-orientations. Les techniques classiques d'extraction de points et de description locale ont ensuite été introduites. L'espace-échelle gaussien est l'unique représentation utilisée dans les travaux existants. Les représentations continues en ondelettes, par leur covariance aux similitudes, apparaissent néanmoins comme des candidates naturelles pour ce problème. Qu'il s'agisse de l'espace-échelle gaussien ou des représentations continues en ondelettes, il est nécessaire de discrétiser le paramètre d'échelle, et celui d'orientation dans le cas de représentations directionnelles, pour implémenter des solutions pratiques. Le chapitre suivant évalue l'influence de la discrétisation de l'espace-échelle gaussien sur la robustesse des caractéristiques extraites. Les différentes répétabilités introduites dans ce chapitre, ainsi que les probabilités de bon appariement sachant une erreur d'estimation, seront utilisées tout au long de la thèse pour mener les évaluations.

Chapitre 2

Détection robuste de points, d'échelles, d'orientations, et description SIFT

Les représentations d'images covariantes aux similitudes permettent de modéliser l'ensemble des images d'une même scène prises sur le même axe optique. Un but de la description locale est d'extraire les mêmes points et les mêmes caractéristiques sur cet ensemble d'images. Le chapitre précédent a montré que les représentations covariantes aux similitudes sont continûment paramétrées en échelle et en orientation. L'objectif de ce chapitre est d'analyser l'effet de la discrétisation de ces paramètres sur la robustesse des caractéristiques extraites (points d'intérêt, échelles, orientations), et sur les descripteurs locaux. Dans un premier temps, les méthodes de référence en détection robuste d'échelle et de points d'intérêt sont évaluées en fonction de la discrétisation en échelle. Ensuite, la même évaluation est conduite pour les méthodes d'extraction robuste d'orientation en fonction de la discrétisation en orientation. Enfin, l'analyse de la sensibilité du descripteur SIFT aux erreurs de localisation des points d'intérêt, et aux erreurs d'estimation en échelle et en orientation, permet de connaître la discrétisation la plus grossière pour une performance de description fixée.

2.1 Robustesse des échelles extraites

Cette section rappelle la nécessité d'une analyse multi-échelles dans un but de description, présente les travaux existants en détection robuste d'échelles, puis évalue l'impact de la discrétisation en échelle de l'espace-échelle gaussien sur la robustesse des échelles extraites.

FIG. 2.1 – Analyse sur quatre octaves dans l'espace-échelle gaussien.

2.1.1 Détermination de la plage des échelles analysées

L'absence d'information a priori sur la taille des objets d'intérêt présents dans la scène impose d'analyser les images sur une plage fixée d'échelles. La représentation multirésolution standard en description est l'espace-échelle gaussien. Selon les travaux, le paramètre d'échelle est soit la variance, soit l'écart-type de la gaussienne. Le choix de l'écart-type permet d'avoir un paramètre cohérent avec le paramètre d'échelle des représentations en ondelettes. Une gaussienne discrétisée uniformément en x et en y, et d'écart-type unité, est un filtre passe-bas de bande-passante égale à la fréquence d'échantillonnage. La convolution d'une image échantillonnée à la fréquence de Nyquist avec cette gaussienne n'introduit donc que des pertes minimes. Une analyse de l'image octave par octave se fait en divisant successivement par deux la bande-passante de la gaussienne. La figure 2.1 représente l'image Lena sur quatre octaves successives. L'hypothèse est faite selon laquelle toutes les images naturelles sont si lisses au-delà de quatre octaves qu'il n'y a plus d'information discriminante à extraire. Cette hypothèse permet de fixer l'intervalle $[s_{min}, s_{max}]$ des échelles analysées à $[1, 8]$. La plage d'échelles communes entre deux images de résolution différente peut être faible. Il est donc nécessaire de décrire les régions extraites non pas sur toute la plage d'échelles $[s_{min}, s_{max}]$, mais à une seule échelle, qui doit être robuste aux changements d'échelles. Une telle échelle s'appelle *l'échelle caractéristique* du point d'intérêt considéré.

2.1.2 Extraction robuste d'échelles

Recherche d'un détecteur générique. Les principaux travaux concernant la détection d'*échelles caractéristiques* ont été menés par Lindeberg [Lin94b, Lin94a]. Pour chacune des primitives de la figure 1.4 (page 33) l'échelle caractéristique peut se définir comme une longueur de transition. Les rampes et les barres, phénomènes transitoires mono-dimensionnels, ont une seule échelle caractéristique. Les coins, les jonctions et les blobs elliptiques ont en revanche deux échelles caractéristiques. L'extraction robuste d'échelles peut donc s'effectuer par la détection simultanée de primitives et de leurs échelles caractéristiques. Ce mode d'extraction est coûteux en temps de calcul, il requiert l'utilisation de plusieurs opérateurs dédiés à chacune des primitives d'intérêt, comme ceux proposés dans [Lin94b] pour la détection de rampes, de barres, de jonctions ou de blobs. Pour une description discriminante, il est nécessaire de détecter plusieurs types de structures. La détection par une série d'opérateurs dédiés à chaque structure d'intérêt étant trop coûteuse, il est souhaitable de chercher un unique opérateur pour la détection des échelles caractéristiques de plusieurs types de structures.

Choix du laplacien normalisé. Si un tel opérateur existe, il est nécessairement isotrope (pour ne pas dépendre de l'orientation de la structure), et à moyenne nulle (pour ne pas dépendre du niveau gris moyen). D'après l'évaluation réalisée dans [MS05], le laplacien est le meilleur candidat. Il est calculé à différentes échelles dans l'espace-échelle gaussien. La réponse maximale à travers les échelles donne l'échelle caractéristique du

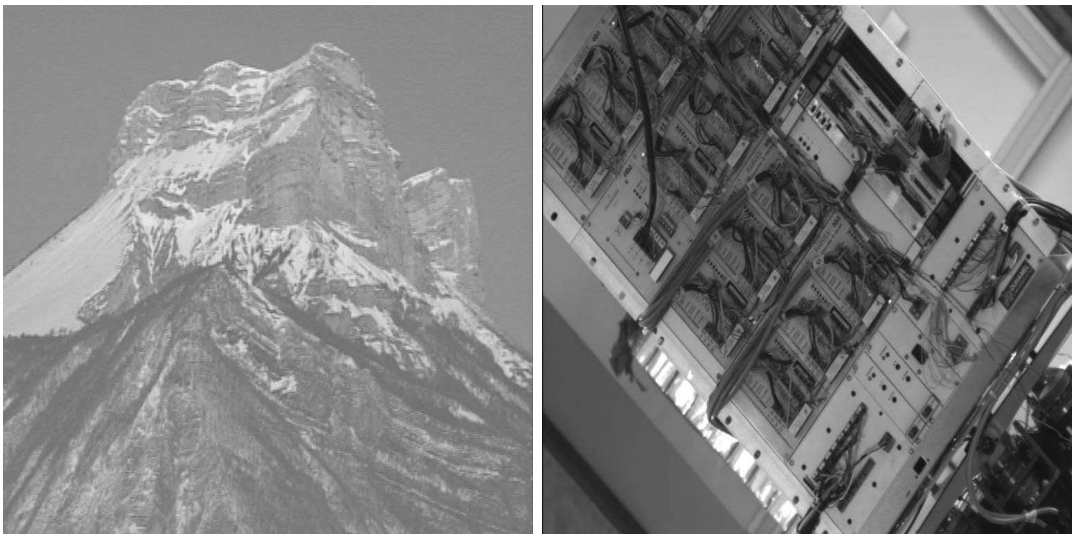

FIG. 2.2 – Images utilisées pour évaluer l'influence de la discrétisation en échelle sur la robustesse des caractéristiques extraites.

point. Le laplacien doit donc être normalisé en échelle. En effet, l'espace-échelle gaussien étant une représentation causale, la dynamique des énergies à différentes échelles n'est pas la même, rendant impossible la comparaison directe. Dans [Lin94b], le laplacien normalisé est défini par :

$$E(\mathbf{x}, s) = s\,|\Delta L(\mathbf{x}, s)| = s\,|L_{xx} + L_{yy}| \tag{2.1}$$

et son maximum local à travers les échelles définit l'*échelle caractéristique* du point considéré.

2.1.3 Impact de la discrétisation en échelle

La recherche des maxima locaux de l'opérateur défini par l'équation 2.1 nécessite de discrétiser en échelle l'espace-échelle gaussien. Compte tenu de considérations psycho-visuelles, l'intervalle $[1, 8]$ des échelles analysées est discrétisé géométriquement en $\{s_n = r^n\}_{0 \leq n \leq \lceil \frac{\log(8)}{\log(r)} \rceil}$. Les précédents travaux [Lin94b, MS05] sur l'extraction robuste d'échelle ne précisent pas l'influence de la raison de cette discrétisation. La raison est choisie empiriquement et fixée à une valeur de l'ordre de 1.2. Pour le problème conjoint de description et de compression, ce choix est essentiel car il détermine la redondance de la représentation multi-échelles. Il est donc nécessaire d'évaluer son impact sur la qualité de la description, et en particulier sur la robustesse des échelles extraites. Pour ce faire, l'extraction est effectuée à partir des images de la figure 2.2 constituant deux cas extrêmes. L'une est une vue d'un paysage éloigné, l'autre est un zoom sur un objet[1]. L'auteur a pris les deux scènes à des focales différentes et a calculé les homographies

[1] Ces images se trouvent publiquement en ligne sur http://www.inrialpes.fr/lear/people/Mikolajczyk.

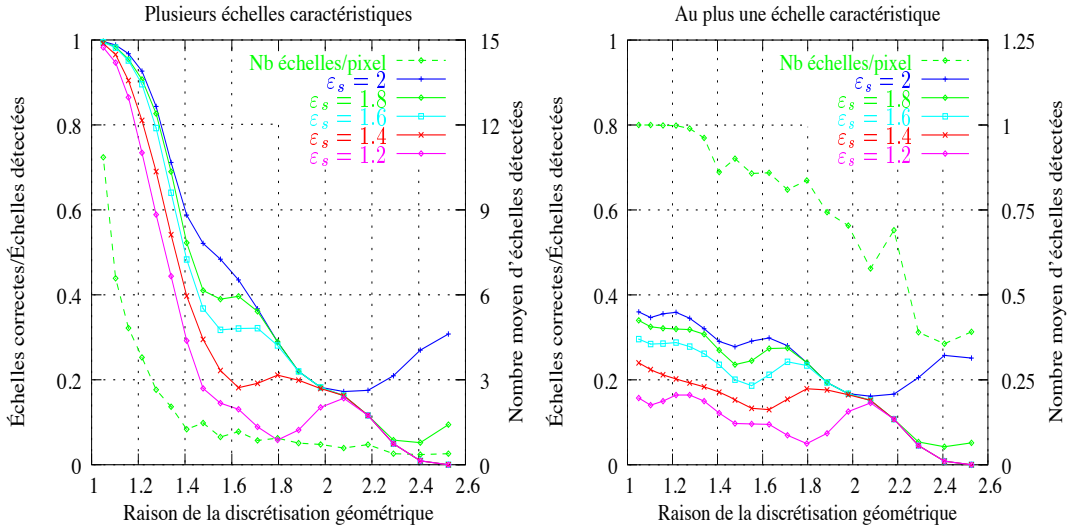

FIG. 2.3 – Influence de la raison de la discrétisation géométrique en échelle sur la robustesse des échelles extraites. En un pixel fixé, tous les maxima locaux à travers les échelles peuvent être extraits (à gauche), ou seulement le plus fort, s'il existe (à droite). Le taux de bonne détection est calculé pour différentes précisions ε_s. Le nombre moyen d'échelles caractéristiques est présenté sur l'axe de droite. Les résultats consistent en la moyenne des détections obtenues à partir d'images des scènes de la figure 2.2 prises à différentes focales.

permettant de faire la correspondance point à point pour chaque paire d'images de la même scène.

Le laplacien normalisé de l'équation 2.1 peut ne pas atteindre de maximum, et dans ce cas aucune échelle n'est détectée, ou au contraire avoir des maxima multiples. Dans ce dernier cas, il est possible d'attribuer plusieurs échelles caractéristiques, ou une seule, celle correspondant au plus fort maximum local.

Se donnant l'erreur d'estimation en échelle ε_s définie par la relation 1.25 (page 44), la figure 2.3 donne le pourcentage d'échelles correctes à ε_s près parmi les échelles détectées selon deux critères différents, et pour différentes valeurs de ε_s. Les échelles sont détectées sur les images de la figure 2.2, pour deux valeurs de focale égales à 1 et 2.2. Les échelles sont détectées en tout point et non pas seulement aux points d'intérêt. Sur la figure de gauche, la détection est considérée comme correcte dès qu'une échelle est correcte parmi toutes celles détectées au point considéré (cas de plusieurs maxima locaux à travers les échelles). Lorsque la raison de la discrétisation géométrique vaut 1.05, il y a, en moyenne, plus de 10 échelles détectées par pixel. Naturellement, la probabilité est grande qu'il existe une échelle correcte parmi ce grand nombre d'échelles détectées. Lorsque la discrétisation en échelle devient de plus en plus grossière, le nombre moyen d'échelles détectées diminue très rapidement, et donc le taux de bonne détection diminue également. La diminution de ce taux est d'autant plus rapide que la précision de bonne détection est fine. Sur la figure de droite, au plus une échelle caractéristique est extraite

en un pixel fixé, celle correspondant au plus fort maximum local à travers les échelles (lorsqu'il existe). Le taux de bonne détection est donc beaucoup plus faible. Le premier point important est que ce taux ne dépend que faiblement de la discrétisation en échelle. À discrétisation fine, il y a beaucoup de maxima locaux à travers les échelles, ce qui pénalise la probabilité d'extraire la bonne échelle. À discrétisation grossière, il n'y a plus qu'un seul maximum local à travers les échelles, et les courbes des figures de gauche et de droite se confondent. Le second point important est la diminution importante du taux de bonne détection avec la précision de bonne détection. Il sera donc nécessaire de concevoir des descripteurs locaux peu sensibles aux erreurs d'estimation en échelle. Un phénomène particulier est le rebond marqué par le taux de bonne détection entre 1.6 et 2 selon la précision de bonne détection. Ceci s'explique par la coïncidence entre la raison de la discrétisation et le facteur de changement de focal égal à 2.2. Pour $\varepsilon_s = 2$, le rebond se trouve en 1.6 car c'est à cette discrétisation que se trouve le plus grand nombre de paires d'échelles correctes à une octave près entre les suites géométriques $\{s_n = 1.6^n\}_{0 \leq n \leq 4}$ et $\{\breve{s}_n = 1.6^n/2.2\}_{0 \leq n \leq 4}$. Lorsque la précision de bonne détection ε_s diminue, le nombre maximal de paires d'échelles correctes entre les deux suites se rapproche de la valeur du facteur de changement de focale, ici égal à 2.2.

Il s'en suit trois observations. La première observation est que le laplacien normalisé ne permet pas de détecter une unique échelle caractéristique en tout point. Il est donc à craindre qu'il soit nécessaire de décrire les points extraits à plusieurs échelles. En fait, la section suivante montrera que la robustesse des échelles caractéristiques est beaucoup plus grande en un point d'intérêt qu'en un point quelconque, et en pratique suffisante. La seconde observation est que les descripteurs devront être assez peu sensibles aux erreurs d'estimation en échelle. La dernière observation est que la sensibilité du taux de bonne détection à la finesse de discrétisation en échelle est relativement. Ce dernier point est encourageant : des représentations multi-échelles faiblement redondantes pourront être utilisées pour traiter le problème conjoint de compression et de description.

2.2 Extraction robuste de points

Dans cette section, les détecteurs de Lindeberg, de Harris, de Harris-Laplace, et de Förstner, présentés dans la section 1.2.1 (page 34), sont évalués en fonction de la discrétisation en échelle de l'espace échelle gaussien.

2.2.1 Détection mono-échelle

Détecteurs de Harris et de Förstner. Les détecteurs mono-échelles de Harris et de Förstner sont tous les deux calculés à partir de la matrice structure définie en 1.19 (page 36). Les dérivées partielles définissant cette matrice sont calculées à une échelle unique, appelée échelle de dérivation, et généralement fixée à une valeur de l'ordre de 1. Les dérivées partielles sont sommées par la convolution avec une gaussienne w_s. L'écart-type de cette gaussienne est l'échelle d'intégration, et est fixée à 1.4. La figure 2.4 donne,

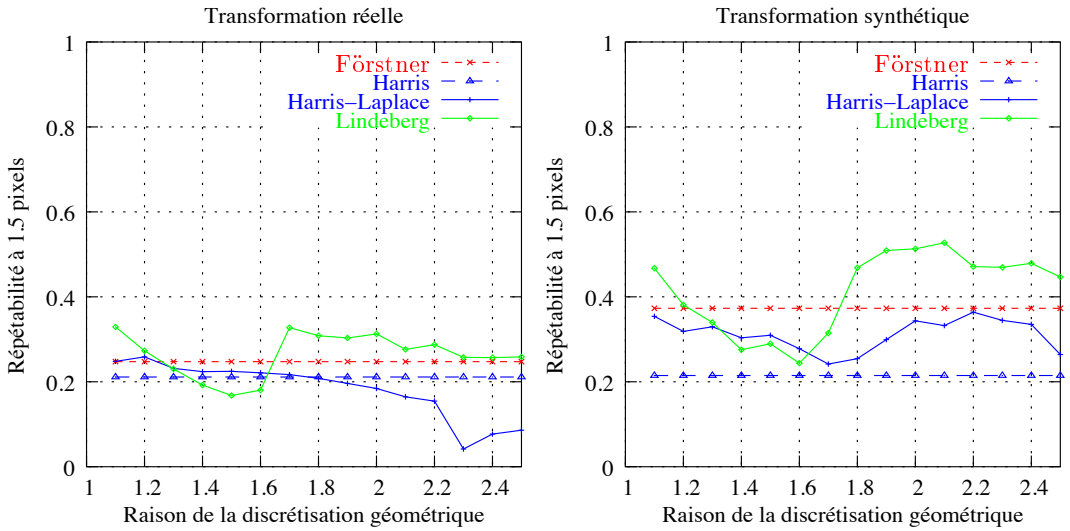

FIG. 2.4 – Répétabilités comparées des détecteurs de Förstner, Harris, Harris-Laplace, et Lindeberg, en fonction de la discrétisation en échelle de l'espace-échelle gaussien. Les détecteurs mono-échelle sont représentés en pointillé, et les détecteurs multi-échelles en trait plein. La répétabilité est calculée pour une précision de 1.5 pixels, pour une transformation réelle (changement de focale de facteur 2.2 sur la figure de gauche), et pour une transformation synthétique (composée d'une dilatation de facteur 2.2, d'une rotation de 30 degrés, d'un crop de 30%, et d'une compression JPEG de facteur 10), sur la figure de droite.

pour ces deux détecteurs, la répétabilité $r_p(1.5)$ telle que définie par la relation 1.27. Cette répétabilité est tracée en pointillé, car elle n'est en fait calculée qu'à une seule échelle. Pour le détecteur de Harris, le paramètre α apparaissant dans la relation 1.20 est fixé à 0.06. Dans le cas d'une transformation réelle mettant en jeu un changement de focale de facteur 2.2, les deux détecteurs conduisent à une répétabilité similaire et de l'ordre de 20%. Pour une transformation synthétique composée d'une dilatation, d'une rotation, d'une compression JPEG, et d'un crop, la répétabilité du détecteur de Förstner est de 40% alors que celle de Harris est de 20%. Le détecteur de Harris extrait non pas seulement des coins et des jonctions, mais aussi des contours, ce qui pénalise sa précision de détection.

Comparaison avec les détecteurs multi-échelles. Le résultat le plus étonnant est que les robustesses des détecteurs de Harris mono et multi échelles sont comparables, et ce, même pour une transformation mettant en jeu un changement de focale ou une dilatation importante. Ce résultat a déjà été observé dans [FB04]. Pour la transformation réelle de la figure de gauche, le pas de discrétisation de l'espace-échelle doit être inférieur à 1.6 pour que le détecteur multi-échelles soit comparable au détecteur mono-échelle. En revanche, pour la transformation synthétique de la figure de droite, le détecteur multi-échelle est meilleur. Même si l'avantage de la détection multi-échelles semble faible

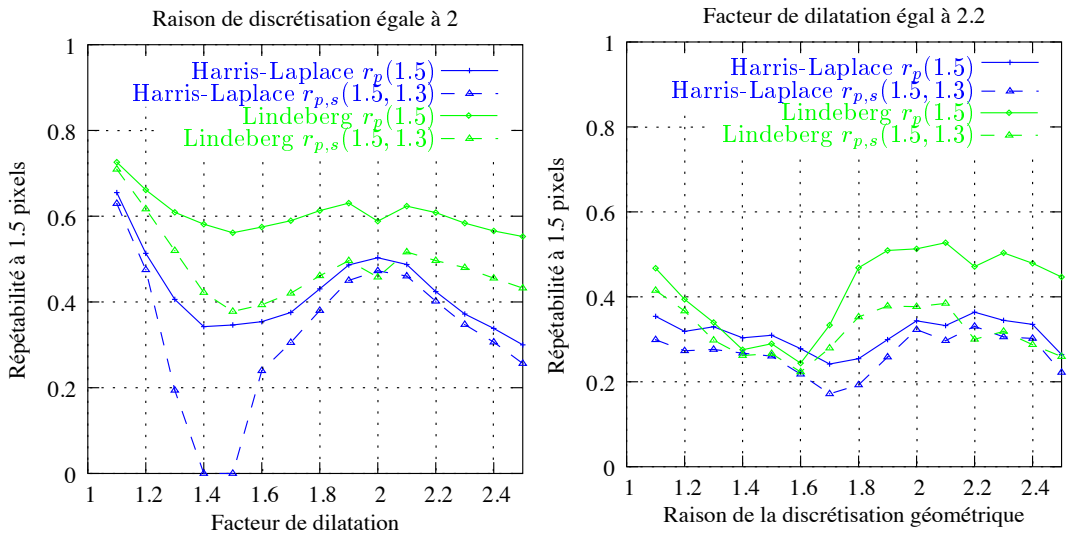

FIG. 2.5 – Évaluation des détecteurs multi-échelles par les répétabilités $r_p(1.5)$ et $r_{p,s}(1.5, 1.3)$ telles que définies par les relations 1.27 et 1.28, page 44. La copie est soit une simple dilatation de l'image originale (figure de gauche), soit composée d'une dilatation, d'une rotation, d'un crop et d'une compression JPEG (figure de droite). Figure de gauche: robustesse des points et de leur échelle caractéristique en fonction du facteur de dilatation, à raison de discrétisation fixée et égale à 2. Figure de droite: Robustesse des points et de leur échelle caractéristique en fonction de la raison de discrétisation, à facteur de dilatation fixée et égale à 2.2.

eu égard de la complexité algorithmique et de la redondance de la représentation, la section 2.5 montrera que, dans une perspective de description, une telle détection est nécessaire. En effet, l'absence d'information sur l'échelle des points extraits, contraint à décrire tous les voisinages à la même échelle, ce qui implique de mauvaises performances de description.

2.2.2 Détection multi-échelles

Cette section évalue en fonction de la discrétisation en échelle les deux détecteurs multi-échelles de points les plus utilisés.

Détecteur de Lindeberg. Les points extraits sont les maxima locaux en espace et en échelle du laplacien normalisé défini par la relation 2.1. Ils sont donc par définition extraits à leur échelle caractéristique. Les maxima locaux sont calculés après discrétisation géométrique en échelle ($s \in \{s_n = r^n\}_n$). La figure 2.4 donne, pour un changement de focale de facteur 2.2 à partir des images de la figure 2.2, la répétabilité à 1.5 pixels, en fonction de la raison r de la discrétisation géométrique en échelle. À discrétisation fine en échelle, le détecteur de Lindeberg dédié aux blobs est le détecteur le plus répétable. Ce détecteur est aussi le détecteur le plus sensible à la discrétisation en échelle, si bien

qu'il devient moins bon que le détecteur de Harris-Laplace pour des pas de discrétisations supérieurs à 1.3. Ce comportement s'observe aussi bien pour la transformation réelle de gauche que pour la transformation synthétique de droite. Pour ces deux types de transformations apparaît un rebond important à partir de la raison égale à 1.6. Pour des discrétisations très grossières, la proportion de points extraits à la première échelle (c'est-à-dire à l'échelle $s_0 = 1$ dans la suite $\{s_n = r^n\}_{0 \leq n \leq \lceil \frac{\log(8)}{\log(r)} \rceil}$) devient importante. Par conséquent, lorsque la raison r coïncide avec le facteur de changement de focale, la détection de points sur l'image originale et sur la copie devient identique, et les points extraits sont très bien localisés. Pour une discrétisation grossière, de raison égale à 2, la figure 2.5 de gauche donne la répétabilité des points d'intérêt en fonction du facteur de dilatation. Le phénomène de rebond permet au détecteur de Lindeberg d'être très répétable sur une large plage de dilatations. La figure de droite permet de constater que, pour une copie composée d'une dilatation de facteur 2.2 et d'autres transformations, les échelles extraites sont très robustes, même à fine précision de bonne détection, fixée à $\varepsilon_s = 1.3$. Pour des raisons inférieures à 1.6, plus de 90% des points d'intérêt ont leur échelle caractéristique correcte. Dans la section 2.1.3, ce taux chute à 15% pour des points quelconques. Le choix de $\varepsilon_s = 1.3$, pour la précision de bonne détection, est dicté par la sensibilité de la description aux erreurs de détection en échelle (se reporter à la section 2.5).

Détecteur de Harris-Laplace. Une méthode est proposée dans [SM96] pour détecter les points de Harris à leur échelle caractéristique. Pour cela, un rapport de proportionnalité égal à 1.5 est fixé entre l'échelle de dérivation et l'échelle d'intégration apparaissant dans la relation 1.19 (page 36). Des points préliminaires sont d'abord extraits. Ce sont les maxima locaux en espace de la mesure de Harris définie en 1.20 à différentes échelles. Les points finalement retenus sont ceux où le laplacien normalisé défini par la relation 2.1 présente un maximum local en échelle. La figure 2.4 montre que la répétabilité de ce détecteur est moins bonne que celle du détecteur de Lindeberg dédié aux blobs. Ceci provient du fait qu'un grand nombre de points préliminaires ne sont pas extraits à leur échelle caractéristique, et sont donc éliminés, alors qu'ils sont très énergétiques, donc répétables. Cet élagage est d'autant plus important que la probabilité est faible qu'un point préliminaire corresponde à un maximum local en échelle du laplacien normalisé. La section 2.5 a montré que cette probabilité diminue fortement avec la raison de la discrétisation en échelle. Par conséquent, on peut observer sur la figure 2.5 de gauche que, pour une discrétisation grossière, la répétabilité des points extraits chute rapidement avec le facteur de dilatation. Cette chute est encore plus forte pour les échelles extraites. Ce détecteur requiert donc une discrétisation très fine en échelle, et n'est donc pas adapté au problème conjoint de compression et de description.

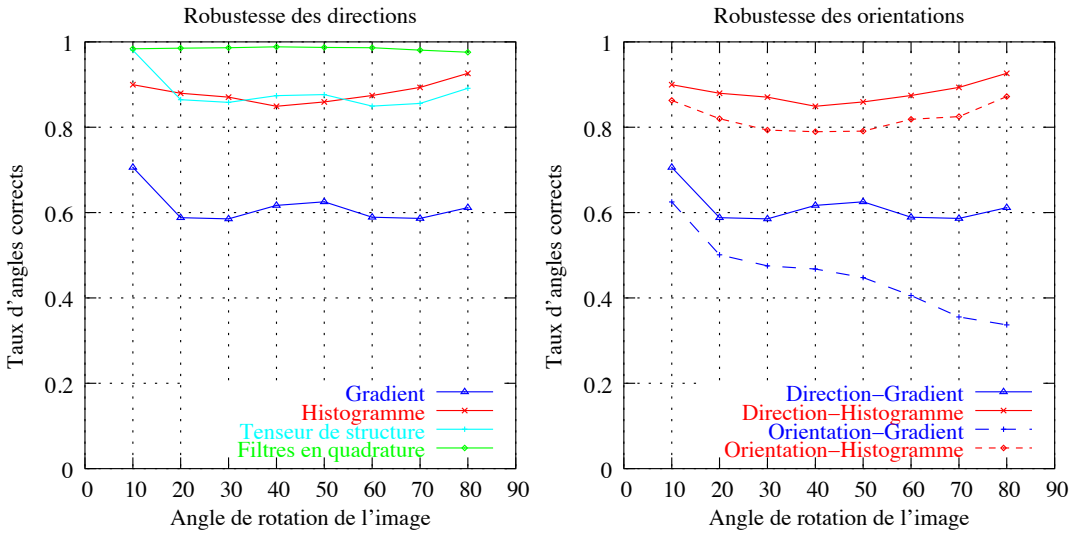

FIG. 2.6 – Taux de directions correctes (figure de gauche) et d'orientations correctes (figure de droite), à 15 degrés près (modulo 180 ou 360 degrés), détectées au niveau des points extraits par le détecteur de Lindeberg. L'évaluation est faite à partir d'images d'une même scène prise à différents angles autour de l'axe optique.

2.3 Extraction robuste d'orientations

L'orientation est une information pertinente pour de nombreux traitements visuels, et en particulier pour la description. Sa définition est toutefois délicate : l'orientation n'a pas de sens au centre de structures isotropes, et peut être multiple sur des coins ou des jonctions. Sur les primitives saillantes illustrées sur la figure 1.4 (page 33), l'orientation peut être définie sur $[0, 2\pi[$ (exemple des rampes) ou seulement sur $[0, \pi[$ (exemple des barres). Dans ce qui suit, on distingue les *orientations* définies sur $[0, 2\pi[$ des *directions* définies sur $[0, \pi[$.

Orientation du gradient. Une première définition de la direction est la direction du gradient égale à :

$$d(\mathbf{x}, s) = arctg(L_x(\mathbf{x}, s)/L_y(\mathbf{x}, s)) \tag{2.2}$$

où L_x, L_y sont les dérivées partielles de l'image I définies par la relation 1.14 (page 32). L'extraction d'orientation s'effectue à l'échelle caractéristique du point, et est donc dépendante de la robustesse des échelles extraites. De même, on peut définir l'orientation par :

$$\begin{aligned} \theta(\mathbf{x}, s) &= atan2((L_x(\mathbf{x}, s), L_y(\mathbf{x}, s)) \\ &= \begin{cases} arctg(L_x(\mathbf{x}, s)/L_y(\mathbf{x}, s)) \text{ si } L_y(\mathbf{x}, s) > 0 \\ \pi + arctg(L_x(\mathbf{x}, s)/L_y(\mathbf{x}, s)) \text{ si } L_y(\mathbf{x}, s) < 0 \text{ et } L_x(\mathbf{x}, s) > 0 \\ -\pi + arctg(L_x(\mathbf{x}, s)/L_y(\mathbf{x}, s)) \text{ si } L_y(\mathbf{x}, s) < 0 \text{ et } L_x(\mathbf{x}, s) < 0 \end{cases} \end{aligned} \tag{2.3}$$

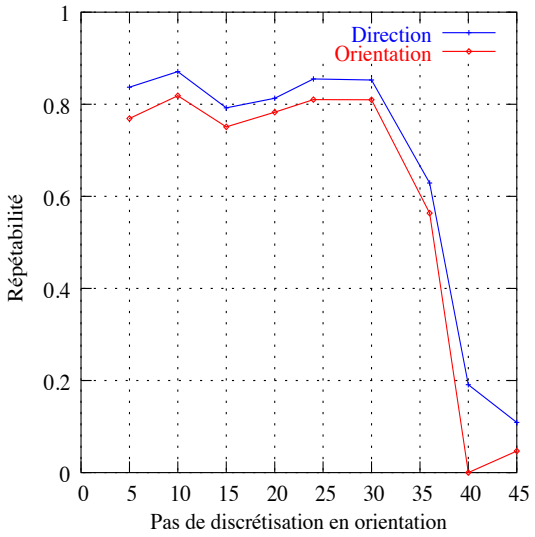

FIG. 2.7 – Influence de la discrétisation en orientation sur la robustesse des orientations extraites. L'évaluation est faite à partir de copies créées par rotation d'angle 60 degrés.

La robustesse des orientations est évaluée à partir d'images d'une même scène prises au même endroit à différents angles autour de l'axe optique. Le pourcentage d'orientations correctes à 15 degrés près est donné par la figure 2.6. Le choix de la précision de 15 degrés est dicté par la description, qui reste robuste pour une telle erreur de détection en orientation. La faible robustesse de cette méthode s'explique par la rapide variabilité du gradient, très importante au niveau des points d'intérêt. Deux idées permettent de limiter cet effet. La première consiste à prendre en compte les orientations voisines, la seconde à estimer l'orientation indépendamment de la phase du signal.

Maximum de l'histogramme des orientations voisines. Cette technique s'applique aussi bien pour les directions que pour les orientations. L'orientation en un point $\mathbf{x}$ dépend de l'ensemble des gradients $\{\nabla L(\mathbf{y}, s)\}_{\mathbf{y} \in V(\mathbf{x})}$ calculés dans le voisinage $V(\mathbf{x})$ à l'échelle s caractéristique du point $\mathbf{x}$. Dans [Low99], les orientations voisines $\theta(\mathbf{y}, s)$, calculées à partir de l'équation 2.3, contribuent à un histogramme d'orientations discrétisées par pas de 10 degrés. Chaque orientation est pondérée par la norme du gradient et par un facteur gaussien d'éloignement au centre $\mathbf{x}$. L'orientation dominante extraite au point $\mathbf{x}$ est celle où l'histogramme des orientations locales atteint son maximum. Cette technique permet de détecter des orientations secondaires, dans les cas où l'histogramme présente des maxima d'amplitudes similaires. La figure 2.6 montre que, pour une rotation de 30 degrés, 75% des points extraits ont leur orientation robuste à 15 degrés près. Il s'agit de la seule méthode robuste permettant de détecter des orientations, et non pas seulement des directions. Le gain en performance de description à utiliser des orientations est important. Cette technique sera donc, dans la suite de la thèse, la technique de référence pour estimer les orientations. Le pas de discrétisation utilisé pour

former l'histogramme des orientations voisines jouera un rôle important dans la redondance des représentations utilisées. La figure 2.7 montre qu'il est possible de discrétiser jusqu'à un pas de 30 degrés tout en conservant une bonne robustesse d'estimation.

Matrice de structure. Les directions voisines peuvent être prises en compte d'une autre manière que par un histogramme. La direction dominante, donnée par le vecteur unitaire $\mathbf{n_p}$, est celle qui minimise la somme des erreurs quadratiques pondérées par un facteur gaussien:

$$\mathbf{n_p} = \arg \min_{\mathbf{n}} \varepsilon_{\mathbf{x}}(\mathbf{n})$$

L'erreur $\varepsilon_{\mathbf{x}}(\mathbf{n})$ est définie par :

$$\varepsilon_{\mathbf{x}}(\mathbf{n}) = \sum_{\mathbf{y}} e^2_{\mathbf{n}}(\mathbf{y})\, w_s(\mathbf{x}-\mathbf{y})$$

où l'erreur $e_{\mathbf{n}}$ est la norme de la projection orthogonale du gradient $\mathbf{g}(\mathbf{y}) = \nabla L(\mathbf{y}, s)$ sur le vecteur unitaire $\mathbf{n}$:

$$e_{\mathbf{n}}(\mathbf{y}) = ||\mathbf{g}(\mathbf{y}) \;-\; ({}^t\mathbf{g}(\mathbf{y}).\mathbf{n})\,\mathbf{n}||$$

Le choix entre $\mathbf{n}$ et $-\mathbf{n}$ étant indifférent, cette technique ne s'applique qu'à la détection de directions et non pas d'orientations. L'erreur quadratique à minimiser s'écrit :

$$\begin{aligned}
\varepsilon_{\mathbf{x}}(\mathbf{n}) &= \sum_{\mathbf{y}} ||\mathbf{g}(\mathbf{y}) \;-\; ({}^t\mathbf{g}(\mathbf{y}).\mathbf{n})\,\mathbf{n}||^2 w_s(\mathbf{x}-\mathbf{y}) \\
&= \sum_{\mathbf{y}} [{}^t\mathbf{g}(\mathbf{y}).\mathbf{g}(\mathbf{y}) \;-\; ({}^t\mathbf{n}.({}^t\mathbf{g}(\mathbf{y}).\mathbf{g}(\mathbf{y})).\mathbf{n})]\; w_s(\mathbf{x}-\mathbf{y}) \\
&= \sum_{\mathbf{y}} [{}^t\mathbf{g}(\mathbf{y}).\mathbf{g}(\mathbf{y})] w_s(\mathbf{x}-\mathbf{y}) \;-\; {}^t\mathbf{n}.\left(\sum_{\mathbf{y}} [{}^t\mathbf{g}(\mathbf{y}).\mathbf{g}(\mathbf{y})] w_s(\mathbf{x}-\mathbf{y})\right).\mathbf{n}
\end{aligned}$$

Minimiser $\varepsilon_{\mathbf{x}}(\mathbf{n})$ revient donc à maximiser $\mathbf{n}^T M(\mathbf{x}, s)\mathbf{n}$, où $M(\mathbf{x}, s)$ est la matrice définie en 1.19 comme la matrice d'autocorrélation des dérivées partielles à l'échelle caractéristique s (aussi appelée tenseur de structure). L'orientation (ou la direction) dominante est donc celle du plus grand vecteur propre de $M(\mathbf{x}, s)$. La figure 2.6 montre que la robustesse des orientation extraites est similaire à celle obtenue par la méthode de l'histogramme pour les rotations inférieures à 30 degrés et devient moins bonne pour les grandes rotations.

Filtres en quadrature. Un filtre en quadrature est un filtre complexe dont la partie réelle est la transformée de Hilbert de la partie imaginaire, c'est-à-dire déphasée de $\frac{\pi}{2}$ dans le domaine de Fourier. Un exemple de filtre en quadrature est le filtre de Gabor orienté verticalement et défini par :

$$h(x, y) = e^{i\alpha x} e^{-\frac{x^2+y^2}{\sigma}} \tag{2.4}$$

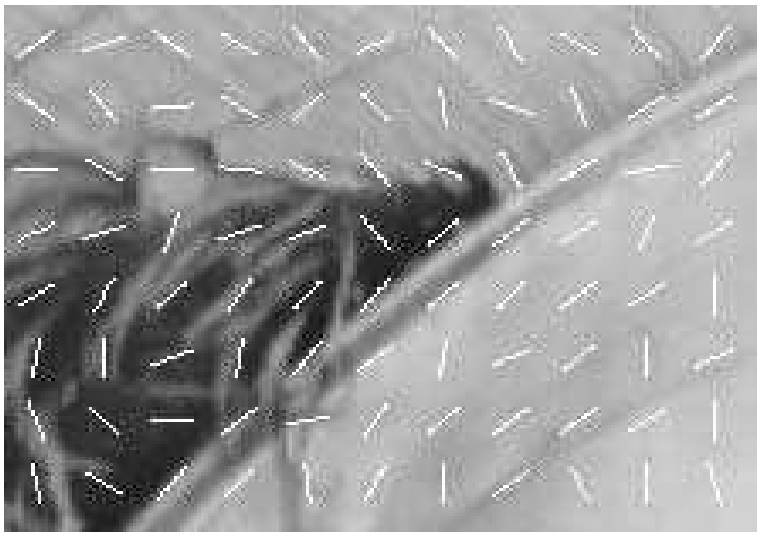

FIG. 2.8 – Exemple de directions extraites sur une portion du chapeau apparaissant dans l'image Lena.

La partie réelle d'un filtre en quadrature orienté verticalement est paire, et permet de détecter les barres, tandis que la partie imaginaire est impaire, et est sensible aux rampes. La magnitude de la réponse obtenue par filtrage avec un filtre en quadrature est indépendante de la phase du signal. Cette propriété est utilisée dans [GK95] pour trouver la direction d'*images simples*, c'est-à-dire d'images de la forme $I(\mathbf{x}) = f(\mathbf{n}^T\mathbf{x})$, où f est une fonction mono-dimensionnelle et $\mathbf{n}$ un vecteur unitaire. Pour ce type d'images, la direction est clairement définie et est donnée par le vecteur $\mathbf{n}$. Le vecteur $\mathbf{z}$ est formé à partir de k filtres en quadrature de direction φ_k :

$$\mathbf{z} = \sum_k q_k \mathbf{m_k} \tag{2.5}$$

où q_k est la magnitude de la sortie complexe du k^e filtre en quadrature, et $\mathbf{m_k} = (\cos 2\varphi_k, \sin 2\varphi_k)^T$. L'orientation de $\mathbf{z}$ permet de connaître la direction de l'image simple, c'est-à-dire la direction du vecteur $\mathbf{n}$ [GK95] :

$$\arg(\mathbf{n}) = \frac{1}{2}\arg(\mathbf{z}) \tag{2.6}$$

Cette technique est la plus robuste, mais elle ne permet d'estimer que des directions et non pas des orientations. La figure 2.6 montre que l'utilisation de trois filtres en quadrature (nombre minimal de filtres pour assurer la relation 2.6) conduit à une excellente robustesse des directions extraites. La figure 2.8 permet de visualiser les directions ainsi extraites à partir d'une portion du chapeau de l'image Lena.

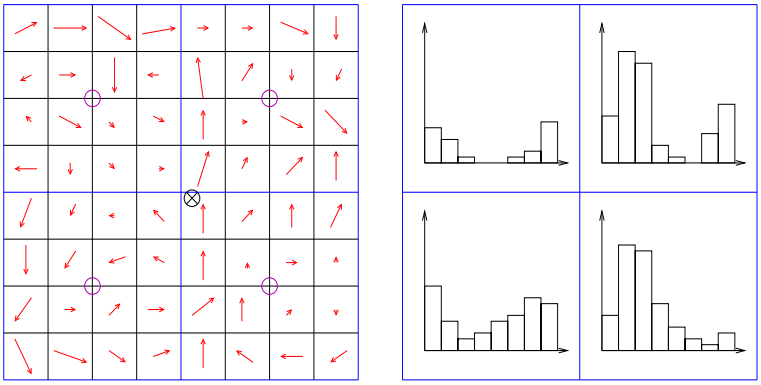

FIG. 2.9 – Description SIFT. Figure de gauche : le point d'intérêt, au centre, est extrait avec une précision sous-pixellique. En chaque point du voisinage de taille 8×8 pixels sont calculées l'orientation (donnée par la direction de la flèche), et l'énergie associée à cette orientation (donnée par la longueur de la flèche). Figure de droite : représentation des histogrammes des orientations calculés dans les 4 voisinages de taille 4×4 pixels, centrés sur les points représentés par des cercles sur la figure de gauche.

2.4 Description SIFT

La description SIFT (Scale Invariant Feature Transform) est une description locale où les points sont extraits à partir d'une représentation par différence de gaussiennes, et décrits à partir des orientations locales. L'évaluation conduite dans [MS05] montre que cette description donne les meilleurs résultats en termes de précision-rappel définie dans la section 1.3.2.

Définition. Les points d'intérêt sont les maxima locaux en espace et en échelle de différences de gaussiennes. L'échantillonnage en échelle est très fin : l'image est sur-échantillonnée par interpolation sur une grille de résolution quatre fois plus fine, puis cinq gaussiennes sont calculées par octave, et la représentation est sous-échantillonnée à chaque nouvelle octave.

Chaque point d'intérêt est décrit selon son voisinage de taille 16×16 à son échelle d'extraction. Dans ce voisinage est calculé l'histogramme des orientations locales. Ces orientations sont simplement estimées par la relation 2.3. Leur contribution à l'histogramme est pondérée par la norme du gradient et par un facteur gaussien d'éloignement au centre du voisinage. L'orientation dominante est l'orientation maximisant l'histogramme discrétisé par pas de 10 degrés. Une orientation secondaire est détectée si un maximum local d'amplitude supérieure à 80% de celle du maximum absolu existe. Pour chaque orientation détectée, un descripteur est calculé à partir des valeurs de la norme et de l'orientation du gradient dans le voisinage 16×16 du point considéré. L'invariance à la rotation est assurée en tournant le voisinage et les orientations qu'il contient d'un angle égal à l'orientation dominante. Cette opération étant effectuée, la figure 2.9

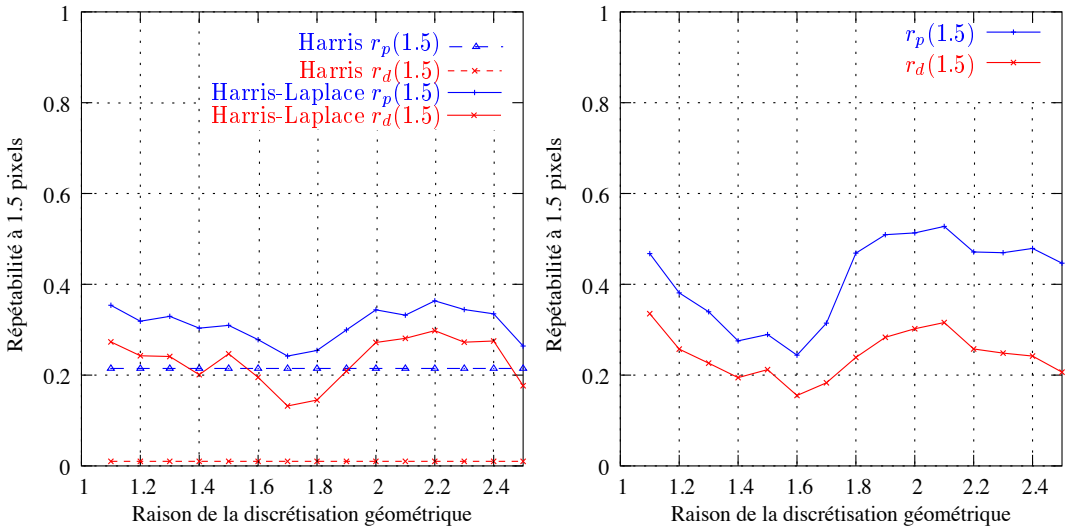

FIG. 2.10 – Influence de la discrétisation en échelle sur la robustesse des descripteurs locaux. Sur la figure de gauche, les points sont extraits par Harris et Harris-Laplace et décrits par les invariants différentiels. Sur la figure de droite, ils sont extraits par le détecteur de Blobs et décrits par SIFT. L'évaluation est faite à partir de copies créées par une dilatation de facteur 2.2, une rotation d'angle 30 degrés, un crop de 30%, et une compression JPEG de facteur 10.

illustre la construction du descripteur SIFT. Par commodité, le voisinage représenté est de taille 8×8 au lieu de 16×16, et est découpé en 2×2 carrés au lieu de 4×4 carrés. Pour chaque carré est construit l'histogramme des orientations locales discrétisé sur 8 valeurs. Le descripteur consiste en la concaténation de ces histogrammes d'orientations. La dimension du descripteur est donc $4 \times 4 \times 8 = 128$. Pour limiter la sensibilité du descripteur aux erreurs de localisation apparaissant lors de l'extraction des points d'intérêt, chaque coefficient du voisinage 16×16 contribue à toutes les dimensions du descripteur. Le poids de la contribution d'un échantillon à une dimension fixée est le produit entre la norme du gradient et un poids linéaire valant $1 - d$, où d est la distance normalisée entre l'échantillon et la valeur centrale de la dimension correspondante. L'invariance aux transformations affines de luminance est assurée par la normalisation du descripteur. Pour approcher l'invariance aux transformations non linéaires de luminance, les dimensions d'intensité trop forte sont seuillées à 0.2, et le descripteur est renormalisé.

Évaluation de la performance de la description. La figure 2.10 permet de comparer, indépendamment de la robustesse des points extraits, la description SIFT et la description par invariants différentiels. Le taux d'appariements corrects définis en 1.32 est d'environ 80% pour la description SIFT et les invariants différentiels. Il devient proche de 100% en utilisant la version originale de SIFT décrite dans [Low99] mettant en jeu un sur-échantillonnage de facteur deux de l'image, une méthode de localisation sous-pixellique des points d'intérêt, et l'élimination des points d'intérêt dont la courbure

du gradient est trop grande. Dans les prochains chapitres, l'idée de description locale à partir de la répartition des orientations dans le voisinage sera transposée dans des représentations multirésolution sous-échantillonnées ne permettant pas ces raffinements. Pour mesurer l'impact du sous-échantillonnage, il est nécessaire d'évaluer la description SIFT sans sur-échantillonnage. La sensibilité de la description SIFT aux erreurs d'estimation en échelle et en orientation est évaluée dans la prochaine section. Elle détermine l'échantillonnage le plus grossier qu'il est possible d'effectuer en échelle et en orientation permettant de conserver la robustesse et le pouvoir discriminant de la description SIFT.

Évaluation de la complexité de la description. Parmi les représentations étudiées dans la suite de la thèse, les représentations d'images permettent d'accèder directement à l'orientation et l'énergie permettant de pondérer la contribution de chaque orientation dans la construction du descripteur SIFT. Si l'on extrait n points d'intérêt à partir d'une image constituée de N pixels, et si l'on dispose de l'information d'énergie et d'orientation en tout point, la complexité de la description SIFT se décompose de la façon suivante :

1. l'extraction des maxima locaux est de complexité égale à $4.N$.
2. l'extraction de l'orientation dominante est de complexité égale à $n.(4\times16\times16+36)$ (dans le voisinage 16×16 : 3 opérations pour le calcul du poids et 1 opération pour le calcul de l'histogramme des orientations) ;
3. l'ajustement des énergies et des orientations sur la nouvelle grille définie par l'orientation dominante est de complexité égale à $n.(2\times6\times256\times+2\times5\times256+256)$ (2 produits entre la matrice 2×2 de rotation et la matrice 256×2 des coordonnées des points, puis interpolation de l'énergie et de l'orientation sur la nouvelle grille, et enfin différence entre les orientations et l'orientation dominante) ;
4. le calcul des 4×4 histogrammes d'orientation est de complexité majorée par $n.(16 \times 16 \times 128 \times 4)$ (3 poids et une somme à calculer pour chaque point du voisinage 16×16, et pour chaque composante du descripteur de longueur 128).

Au final, la quasi totatilité du temps de calcul est consacrée à la dernière étape. La complexité de la description SIFT est donc proche de $4N+2.10^5n$ (en supposant l'énergie et l'orientation connues en tout point).

2.5 Sensibilité de la description aux erreurs de détection

La description locale s'effectue en un pixel, une échelle, et éventuellement une orientation, fixés durant la phase de détection de points d'intérêt. Cette détection s'effectue dans l'espace-échelle gaussien discrétisé en échelle. La section précédente a permis d'évaluer l'impact de la discrétisation en échelle sur la robustesse des points d'intérêt et de leurs échelles caractéristiques. Dans ces expériences, une paire de points était dite robuste si les points étaient distants de moins de 1.5 pixels ; une paire d'échelles (s_1, s_2) si $\varepsilon_s = |\log \frac{s_1}{s_2}| \leq \log 1.3$; une paire d'orientations si leur écart était inférieur à 15 degrés.

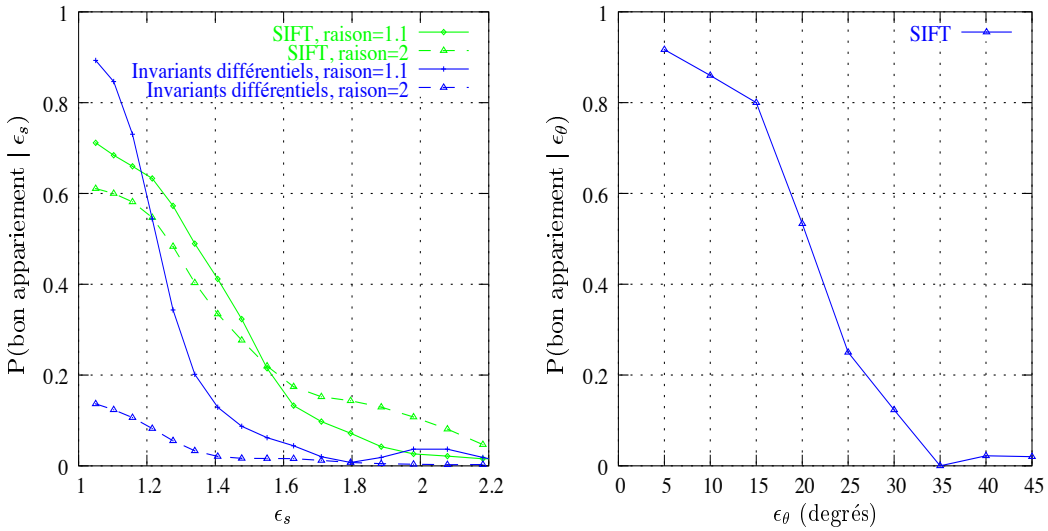

FIG. 2.11 – Sensibilité de la description aux erreurs d'estimation en échelle et en orientation.

Les précisions permettant d'évaluer la robustesse sont fixées par la sensibilité des descripteurs aux erreurs de localisation des points, et aux erreurs d'estimation des échelles et des orientations. Cette section a pour objet l'étude de cette sensibilité. Dans toutes les expériences suivantes, un appariement est dit correct si le descripteur de l'image transformé a pour descripteur le plus proche parmi les descripteurs de l'image originale le descripteur correspondant à un point distant de moins de 1.5 pixels (sachant qu'un tel point existe).

Sensibilité aux erreurs d'estimation en échelle La figure 2.11 de gauche donne la probabilité de bon appariement par descripteurs sachant l'erreur d'estimation en échelle donnée en abscisse (toute autre erreur négligeable par ailleurs). L'erreur ε_s est définie par la relation 1.25 (page 44). À discrétisation géométrique fine en échelle, de raison égale à 1.1, les invariants différentiels sont plus sensibles que les descripteurs SIFT aux erreurs d'estimation en échelle. Malheureusement, cette sensibilité est importante ; la probabilité de bon appariement diminue rapidement avec l'erreur d'estimation en échelle. À discrétisation grossière, les invariants différentiels ne sont plus robustes, alors que les descripteurs SIFT le restent et conservent approximativement la même sensibilité aux erreurs d'estimation. Pour $\varepsilon_s = 1.5$, la probabilité de bon appariement est égale à 0.3. Ce résultat n'est pas encourageant pour transposer les techniques de description dans des représentations à échantillonnage dyadique en échelle. C'est pourtant ce type de représentations qui est utilisé en compression. On verra dans le troisième chapitre qu'une telle transposition est en fait possible, car l'erreur d'estimation en échelle est rarement supérieure à 1.5.

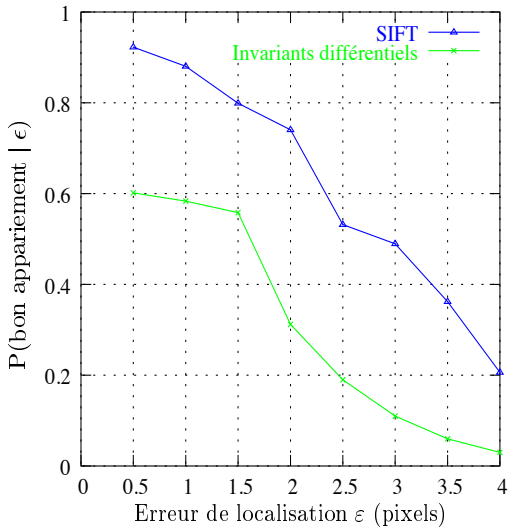

FIG. 2.12 – Sensibilité de la description aux erreurs de localisation des points d'intérêt.

Sensibilité aux erreurs d'estimation en orientation La description SIFT est également très sensible aux erreurs d'estimation en orientation. La figure 2.11 de droite montre que, pour une erreur de 15 degrés, la probabilité de bon appariement par descripteur le plus proche est de 0.8, et devient inférieure à 0.3 pour des erreur supérieures à 25 degrés. Cette sensibilité est naturelle car le descripteur SIFT consiste précisément en la description de la répartition des orientations locales. Les techniques vues dans la section 2.3 permettent une estimation à 15 degrés près, ce qui est suffisant.

Sensibilité aux erreurs de localisation des points d'intérêt La figure 2.12 donne la sensibilité de la description SIFT et des invariants différentiels aux erreurs de localisation des points d'intérêt. Les descripteurs SIFT sont assez peu sensibles, et les invariants différentiels très peu sensibles, aux erreurs de localisation inférieures à 1.5 pixels. C'est la précision choisie dans toute la thèse pour évaluer la répétabilité des points d'intérêt. Pour de plus grandes erreurs de localisation, les dérivées partielles sont trop modifiées pour assurer la stabilité des invariants différentiels. De même, lors du calcul du descripteur SIFT, la partition du voisinage en carrés de taille 4×4 pixels ne tolère pas de grosses erreurs de localisation, même cet effet est limité par la pondération des orientations.

2.6 Conclusion

Les contraintes de linéarité, de covariance aux translations, aux rotations, et aux dilatations, requises pour la description, restreignent les transformées aux convolutions sur des fonctions de base générées par translation, rotation et dilatation. En pratique, les représentations sont échantillonnées en espace, en orientation et en échelle, et par conséquent ne peuvent pas être strictement covariantes. Des expériences ont permis

d'évaluer l'influence de la fréquence d'échantillonnage en échelle et en orientation sur la robustesse des échelles, des points, et des orientations. Il apparaît que l'échantillonnage dyadique en échelle couramment utilisé en compression est suffisant pour la description. Une précision de 15 degrés sur les orientations extraites est nécessaire. Le quatrième chapitre montrera que certaines représentations permettent d'obtenir une telle précision d'estimation en orientation. Le problème de l'échantillonnage en espace est sans doute le plus contraignant pour le problème conjoint de compression et de description. Le prochain chapitre montrera qu'une description robuste est difficile à obtenir à partir de représentations multirésolution à échantillonnage critique (de redondance égale à un). Des représentations à échantillonnage spatial moins fort (donc redondantes) sont étudiées dans le quatrième chapitre dans une perspective de description locale.

Chapitre 3

Description à partir de représentations par ondelettes à échantillonnage critique

La description vise à extraire l'information utile sur l'image pour une application donnée comme la classification, la détection de copies, la reconnaissance d'objets, ou le suivi de cibles. Cette information utile est présente dans les images compressées avec peu de perte. Le problème est de savoir dans quelle mesure le temps de décompression peut être épargné pour extraire cette information. Un schéma classique de compression est composé d'une transformation d'images, d'une quantification, et du codage des coefficients quantifiés. L'inversion de cette dernière étape est inévitable. Les techniques actuelles de codage ne permettent pas de décrire le signal codé. Il s'agit d'une limitation importante puisque le décodage est souvent l'étape la plus coûteuse dans le temps de décompression (c'est par exemple le cas dans le standard JPEG 2000). Le problème de la recherche de techniques de codage adaptées à la description n'est pas traité dans cette thèse. Dans la conception d'un schéma de compression dont l'objectif prioritaire est l'accès rapide à l'information visuelle, la phase de décodage doit être de faible complexité, au prix d'une augmentation du débit. Dans cette thèse, il est supposé que les coefficients quantifiés sont codés par un codeur arithmétique de complexité linéaire, et permettant d'atteindre un débit proche de l'entropie marginale de ces coefficients. La complexité linéaire permet de rendre attractive la description dans le domaine transformé par rapport aux schémas actuels requérant une décompression complète de forte complexité.

Dans un tel schéma de compression adapté à la description locale, le problème central est celui de la représentation des images. Les coefficients transformés puis quantifiés doivent être de faible entropie et permettre l'extraction rapide de l'information utile pour la description. La transformée en ondelettes est une candidate naturelle pour ce problème. Elle permet, d'une part, d'excellents résultats en compression et, d'autre part, de transposer certaines techniques de description grâce à son analyse localisée en espace

et en échelle. Ce chapitre présente d'abord les schémas de compression reposant sur une transformée en ondelettes à échantillonnage critique, il montre ensuite la variance de cette transformée aux translations et aux rotations, et propose, malgré cette variance, un extracteur de points dans le domaine compressé. Enfin, le problème de l'estimation robuste d'orientation est discuté, mais aucune solution satisfaisante n'est proposée.

3.1 Schémas de compression basés ondelettes

Un schéma de compression vise à extraire le *débit* minimum d'information permettant de reconstruire l'image à *distorsion* fixée. Pour y parvenir, les schémas classiques de compression sont composés d'une étape de changement de représentations d'images, d'une étape de quantification, et d'une étape de codage. Cette section introduit les caractéristiques des images naturelles qui permettent de définir des transformées pertinentes pour la compression, c'est-à-dire réversibles et de faible entropie. Une attention particulière est portée sur la manière classique dont est discrétisée la transformée continue pour la rendre adaptée à ce problème. C'est, en effet, cette discrétisation qui est responsable de sa forte variance aux translations et aux rotations, et donc de l'impossibilité de décrire dans le domaine transformé.

3.1.1 Spécificité des images naturelles

Lors de l'acquisition numérique, une image en niveaux de gris de résolution 512×512 quantifiée sur 8 bits coûte 256 Koctets. Cette image est une *image naturelle*, c'est-à-dire une image d'une scène du monde réel. Cela constitue une information a priori permettant de réduire considérablement le coût de représentation. En effet, la probabilité de générer aléatoirement une image ressemblant à une image naturelle est extrêmement faible. L'ensemble des images naturelles est très petit dans l'ensemble des images possibles. Une caractérisation, même partielle, de cet ensemble a de fortes répercussions en débruitage, en compression, ou sur toute tâche visuelle. Il est malheureusement vain de chercher à caractériser cet ensemble par l'analyse de la distribution empirique d'une grande base d'images. Le nombre d'images nécessaire pour estimer cette distribution est beaucoup trop grand, il croît exponentiellement avec la dimension de l'espace image, égale au nombre de pixels. Deux hypothèses sur les statistiques des images naturelles sont couramment utilisées. La première est une hypothèse de Markov selon laquelle la probabilité d'un niveau de gris sachant les valeurs de niveaux de gris dans un petit voisinage est indépendante des valeurs à l'extérieur de ce voisinage. La seconde est une hypothèse de stationnarité du processus de génération des images naturelles. En réalité, les lois des niveaux de gris varient dans l'espace, et le processus n'est pas stationnaire. Cette hypothèse est faite par confort théorique, elle permet de dériver la représentation d'images optimale en compression.

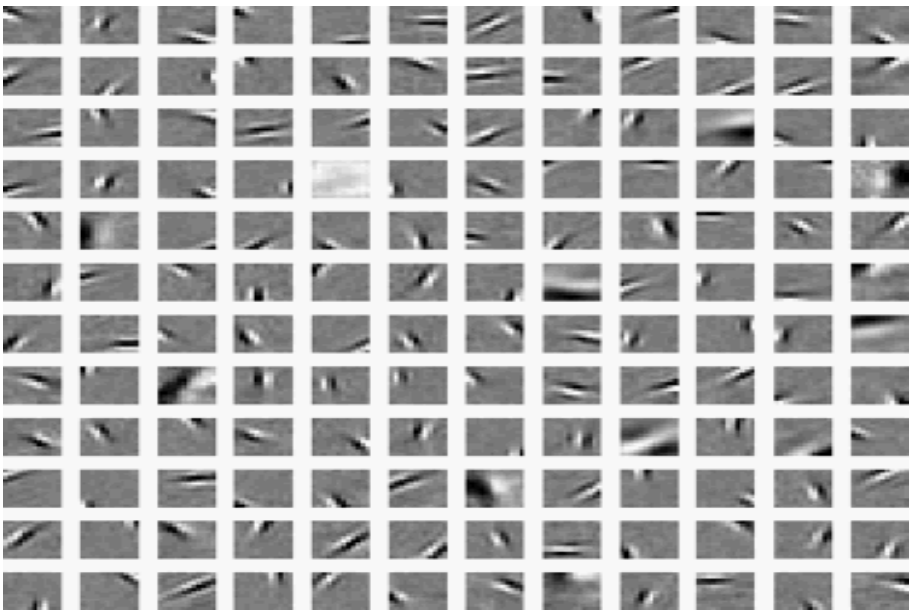

FIG. 3.1 – Filtres générés par ACI à partir d'un grand ensemble d'images de taille 12×12 pixels (extrait de [OF97]).

ACP et TCD. La caractéristique la plus évidente des images naturelles est la forte corrélation entre pixels proches. Sous hypothèse de stationnarité, la matrice d'auto-corrélation est circulante et donc diagonalisable par une base de Fourier. Cette analyse en composantes principales (ACP) est optimale en compression, en ce sens que le sous-espace engendré par les n plus grands vecteurs propres est le sous-espace de dimension n minimisant l'erreur quadratique entre l'image originale et sa projection linéaire. La transformée en cosinus discrète (TCD) est une bonne approximation de l'ACP sur base fixe, et est utilisée en compression dans le format JPEG. L'hypothèse d'auto-similarité des statistiques à travers les échelles contraint les coefficients de l'ACP (et de la TCD) à suivre une certaine loi. En effet, cette hypothèse n'est satisfaite que si l'espérance des spectres de puissance décroît avec une puissance de la fréquence [MG01]. Dans les années 1950 avait déjà été constatée une décroissance statistique avec le carré de la fréquence. Cette rapide décroissance permet de reconstruire une image de bonne qualité avec peu de coefficients de la TCD, rendant possible la compression JPEG.

ACI. Pour obtenir une meilleure caractérisation, les statistiques d'ordre supérieur à deux doivent être prises en compte. Le filtrage des images naturelles par des filtres passe-bande conduit à des distributions fortement non gaussiennes [BA83, Fie87, Mal89], largement piquées à l'origine et à queue allongée. Ces distributions proviennent de la structure des images naturelles composées de régions lisses séparées par des contours. Les régions lisses contribuent aux petites amplitudes formant le pic à l'origine, et les contours aux fortes amplitudes en queue de distribution. Les coefficients de la TCD sont

décorrélés mais non indépendants, le processus de génération d'images naturelles étant non gaussien. L'analyse en composantes indépendantes (ACI) repose sur l'optimisation d'une mesure de non gaussianité comme le kurtosis, égal au moment d'ordre quatre divisé par le carré de la variance. Cette analyse, effectuée au milieu des années 1990 dans [BS97, OF97] à partir d'un grand ensemble d'images de taille 12×12 pixels, a conduit aux filtres présentés dans la figure 3.1.

Ondelettes. Les composantes indépendantes ainsi définies sont des filtres passe-bande orientés de bande-passante voisine d'une octave. Elles ressemblent fortement aux ondelettes définies sur grille dyadique et présentées dans la prochaine section. Cette découverte a permis des progrès considérables en compression pour deux raisons. D'une part, l'entropie des distributions des coefficients d'ondelettes est beaucoup plus faible que celle des coefficients de la TCD. D'autre part, la dépendance entre coefficients d'ondelettes peut être prise en compte plus efficacement que celle entre coefficients de la TCD. Contrairement à ce que son nom indique, les coefficients obtenus par ACI ne sont pas indépendants. De même, les coefficients d'ondelettes sont décorrélés mais non indépendants. Les coefficients adjacents en position, échelle ou orientation, sont fortement corrélés en valeur absolue. Shapiro [Sha93] a trouvé une heuristique permettant de prendre suffisamment bien en compte les dépendances inter-échelles pour obtenir un coût de codage inférieur à la limite entropique fixée par la distribution marginale. Les efforts se portent aujourd'hui sur l'élaboration d'ondelettes anisotropes aptes à saisir la géométrie des contours présents dans les images naturelles [Kin98, Gop03, CD99, DV00, Do01, VBLVD, FA91], et sur la recherche de dictionnaires redondants permettant une représentation creuse des images naturelles [MZ93].

3.1.2 Transformées continues en ondelettes

Jusqu'à une époque récente, les transformées de Fourier et de Gabor étaient les seules alternatives à la représentation des images en niveaux de gris. Les décompositions pyramidales et les transformées en ondelettes ont donné naissance à de nombreuses autres représentations. La nature discrète des images conduit également à une multiplication des représentations possibles : il existe de nombreuses façons de discrétiser une transformée continue. Le but de cette section est d'analyser l'adaptation des transformées continues au problème conjoint de compression et de description.

Transformée de Fourier. Dans cette section, une image I est un élément de $L_2(\mathbb{R}^2)$. Sa *transformée de Fourier* est définie par :

$$\forall \omega \in \mathbb{R}^2, \qquad \hat{I}(\omega) = \int_{\mathbb{R}^2} I(\mathbf{x}) e^{-2i\pi\omega^T \mathbf{x}} \mathrm{d}\mathbf{x} \tag{3.1}$$

La transformée de Fourier $\hat{I}$, aussi appelée spectre de I, est la projection de I sur les modulations pures $e^{2i\pi\omega}$ pour $\omega \in \mathbb{R}^2$. Elle est linéaire, continue, bijective, et isométrique (conserve les normes). Les images naturelles ont la particularité d'avoir un spectre

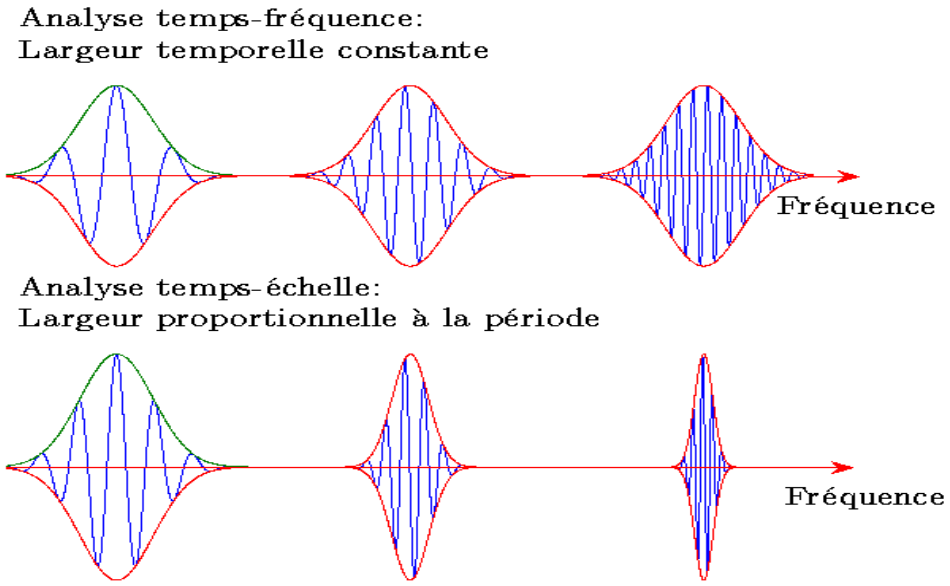

FIG. 3.2 – Fenêtres analysantes dans une analyse espace-fréquence, et dans une analyse espace-échelle.

d'énergie qui suit en espérance une décroissance avec le carré de la fréquence :

$$\mathbb{E}(||\hat{I}(\omega)||^2) = \frac{A}{||\omega||^2} \tag{3.2}$$

L'énergie spectrale est donc concentrée dans les basses fréquences. Il en va de même pour les coefficients de la transformée en cosinus discrète, ce qui la rend intéressante en compression.

Transformée de Gabor. La transformée de Fourier n'est en revanche pas du tout adaptée à la description locale : chaque point du spectre est calculé à partir de toute l'image. En particulier, il est impossible de retrouver la position d'une discontinuité à partir du spectre. Une solution consiste à calculer la transformée sur l'image convoluée par une fenêtre localisée et d'effectuer cette opération pour toutes les positions possibles. C'est le principe de la transformée de Fourier à fenêtre glissante, aussi appelée *transformée de Gabor*. Elle est définie par la projection sur l'ensemble des fenêtres $h_{\mathbf{u},\omega}(\mathbf{x}) = h(\mathbf{x}-\mathbf{u})e^{2i\pi\omega^T\mathbf{x}}$:

$$\forall(\mathbf{u},\omega) \in \mathbb{R}^2\times\mathbb{R}^2, \qquad TG[I](\mathbf{u},\omega) = \int_{\mathbb{R}^2} I(\mathbf{x})\overline{h_{\omega,\mathbf{u}}}(\mathbf{x})\mathrm{d}\mathbf{x} \tag{3.3}$$

où « la fenêtre mère »$h \in L_2(\mathbb{R}^2)$ est centrée à l'origine et d'énergie unité. Tout comme la transformée de Fourier, la transformée de Gabor est linéaire, continue, bijective, et isométrique. Elle réalise une *analyse espace-fréquence* de l'image I : la modulation de la fenêtre d'analyse a pour effet de translater la région spectrale analysée. Une analyse

espace-fréquence est illustrée par la figure 3.2, où la largeur spatiale de la fenêtre Δ_x, définie en 1.15, est la même pour toutes les fréquences analysées. Pour la description locale ou d'autres analyses de phénomènes physiques, il est judicieux d'adapter la largeur spatiale à la fréquence analysée. Lorsque cette adaptation permet d'obtenir un nombre constant d'oscillations dans l'enveloppe analysante, on parle d'*analyse espace-échelle*. Il suffit pour cela de rendre constant le produit $\omega_0 \Delta_x$, où ω_0 est la fréquence centrale de la fenêtre. La résolution fréquentielle relative $\frac{\Delta_\omega}{\omega_0}$ est la même pour toutes les fenêtres d'analyse. L'effet d'une dilatation spatiale de la fenêtre étant une contraction spectrale de même facteur, une analyse espace-échelle est la projection sur la famille de fonctions générées par dilatation et translation :

$$\forall \mathbf{u} \in \mathbb{R}^2,\ s \in \mathbb{R}^{*+}, \qquad \psi_{\mathbf{u},s}(\mathbf{x}) = \frac{1}{s}\psi(\frac{\mathbf{x}-\mathbf{u}}{s}) \tag{3.4}$$

où s est le facteur de dilatation, et $\frac{1}{s}$ un facteur de normalisation permettant de préserver à toute échelle l'énergie de la fenêtre. La figure 3.2 permet de comparer les fenêtres d'analyse utilisées en analyse espace-fréquence et celles utilisées en analyse espace-échelle.

Transformée en ondelettes. Dans une analyse espace-échelle, si « la fenêtre mère »ψ est une ondelette isotrope, la projection de l'image I sur toutes les fenêtres translatées et dilatées s'appelle la transformée de I par l'ondelette ψ. La condition d'admissibilité pour que la fenêtre complexe $\psi \in L_1(\mathbb{R}^2) \cup L_2(\mathbb{R}^2)$ soit une ondelette est :

$$c_\psi = \int_{\mathbb{R}^2} |\hat{\psi}(\mathbf{x})|^2 \frac{\mathrm{d}^2\mathbf{x}}{||\mathbf{x}||^2} \ < \ \infty \tag{3.5}$$

Une condition nécessaire pour que c_ψ soit fini est que $\hat{\psi}$ soit nulle à l'origine, c'est-à-dire que ψ soit de moyenne nulle. Cette condition est suffisante dès que la fonction ψ est suffisamment régulière.

Dans le cas général où l'ondelette n'est pas isotrope, la transformée en ondelettes de l'image I est définie par :

$$\forall (\mathbf{u}, s, \theta) \in \mathbb{R}^2 \times \mathbb{R}^{*+} \times [0, 2\pi[, \qquad T[I](\mathbf{u}, s, \theta) = \frac{1}{s}\int_{\mathbb{R}^2} I(\mathbf{x})\overline{\psi}(s^{-1} r_{-\theta}(\mathbf{x}-\mathbf{u}))\,\mathrm{d}^2\mathbf{x} \tag{3.6}$$

où r_θ est la matrice de rotation d'angle θ. Lorsque l'ondelette ψ est isotrope, il n'y a pas de dépendance en θ. La transformée en ondelettes est linéaire, continue, isométrique, et covariante aux similitudes du plan. La condition d'admissibilité permet d'obtenir la formule de reconstruction :

$$\forall \mathbf{x} \in \mathbb{R}^2, \qquad I(\mathbf{x}) = c_\psi^{-1} \int_{\mathbb{R}^2}\int_{\mathbb{R}^{*+}}\int_0^{2\pi} T[I](\mathbf{u}, s, \theta))\psi(s^{-1} r_{-\theta}(\mathbf{x}))s^{-3}\,\mathrm{d}^2\mathbf{u}\,\mathrm{d}s\,\mathrm{d}\theta \tag{3.7}$$

Comme la transformée est isométrique, la quantité $||T[I](\mathbf{u}, s, \theta)||^2$ est une densité d'énergie dans le domaine transformé. Cette densité possède deux caractéristiques intéressantes pour la compression et la description. D'une part, une ondelette ψ admettant

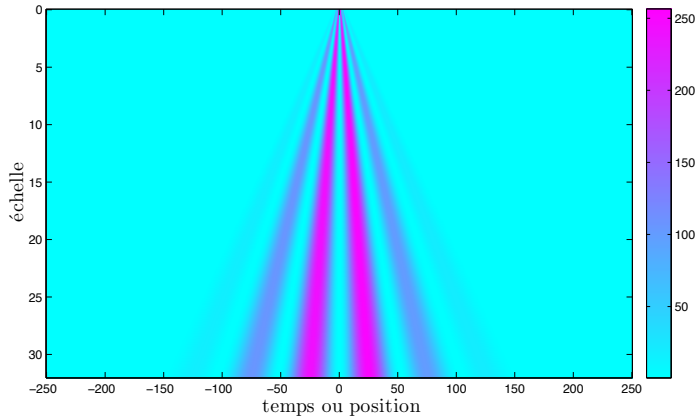

FIG. 3.3 – Densité d'énergie dans l'espace-échelle de l'ondelette de Morlet d'une fonction porte. Les coefficients non nuls sont dans le cône d'influence de la singularité située à l'origine.

$n+1$ moments nuls est orthogonale à tous les polynômes de degré inférieur ou égal à n. Les images naturelles pouvant être approximées par des fonctions polynomiales par morceaux, leur transformée en ondelettes est *creuse* ou *parcimonieuse*. D'autre part, la transformée en ondelettes est un outil privilégié pour détecter les points saillants, utiles pour amorcer le processus de description locale. La figure 3.3 montre la densité d'énergie dans l'espace-échelle d'une singularité modélisée par une fonction porte. Pour une singularité en deux dimensions localisée en (x_0, y_0), l'ensemble des coefficients d'ondelettes affectés forment un cône dans l'espace-échelle défini par l'ensemble des points $(\mathbf{x}, s, \theta)$ tels que :

$$|\mathbf{u}_\theta{}^T \mathbf{x} - x_0| < sC_x$$
$$|\mathbf{u}_{\theta+\frac{\pi}{2}}{}^T \mathbf{x} - y_0| < sC_y$$

où $\mathbf{u}_\theta$ est le vecteur unitaire d'orientation θ, et où le support de l'ondelette est compact et défini par $[-C_x, C_x]\times[-C_y, C_y]$. Ce cône est appelé le *cône d'influence* de la singularité localisée en (x_0, y_0). Pour décrire une singularité, l'information utile est présente dans son cône d'influence.

3.1.3 Transformée discrète en ondelettes

La transformée continue en ondelettes doit être discrétisée pour être applicable en compression. Dans une perspective de compression et de description, les caractéristiques souhaitées de la transformée discrète sont mentionnées ci-dessous.

– Le nombre de coefficients transformés doit être faible. Dans le cas d'une image continue, la connaissance de la transformée $T[I]$ sur $\mathbb{R}^2\times\mathbb{R}^{*+}\times[0, 2\pi[$ tout entier n'est pas nécessaire pour reconstruire l'image I. La reconstruction peut s'effectuer

à partir des coefficients d'ondelettes calculés sur $\{(m2^p, n2^p, 2^p)\}_{m,n,p\in\mathbb{Z}}$ [Mal98]. Dans le cas d'une image discrète, le nombre minimal de coefficients est le nombre de pixels de l'image. On parle dans ce cas de transformée en ondelettes à échantillonnage critique.

- Les coefficients transformés doivent être indépendants. Les niveaux de gris d'une image naturelle sont fortement dépendants, et on ne connaît pas de transformée à coefficients indépendants. Il est revanche possible d'obtenir des coefficients décorrélés par projection sur une base orthogonale de l'espace transformé. L'analyse multirésolution, introduite par Mallat dans [Mal89], et présentée dans cette section, permet de construire des bases orthonormées d'ondelettes.
- Les coefficients transformés doivent être d'entropie marginale, et si possible d'ordre supérieur, faible. La transformée en ondelettes absorbe tous les polynômes de degré inférieur ou égal à la régularité de l'ondelette mère. Par conséquent, la transformée en ondelette est creuse, et donc de faible entropie marginale.
- Il doit exister une implémentation de la transformée discrète de faible complexité. L'analyse multirésolution montre qu'une implémentation de complexité linéaire existe par un banc de filtres a deux canaux.
- La discrétisation de la transformée continue doit affecter aussi peu que possible ses propriétés de de covariance. La section 3.2 montrera que le banc de filtres obtenu par l'analyse multirésolution ne permet pas d'assurer cette caractéristique.
- Les filtres d'analyse doivent être à phase linéaire, donc symétriques. Cette caractéristique est essentielle dans une perspective de description. Les éléments saillants, comme les contours, ont un riche contenu spectral. Pour détecter une primitive saillante particulière, toutes ses fréquences caractéristiques doivent rester en phase (sinon elles se brouillent avec les éléments adjacents).

Le formalisme de l'analyse multirésolution est donc essentiel pour comprendre pourquoi les transformées discrètes utilisées dans le standard JPEG 2000 ne sont pas adaptées au problème conjoint de compression et de description. Dans ce formalisme, un signal $f \in L_2(\mathbb{R})$ se décompose en une bande d'approximation du signal f à l'échelle $p \in \mathbb{N}^*$, et des bandes de détails du signal f aux échelles $0 \leq k \leq p$. Le banc de filtres permettant d'obtenir les composantes d'approximation et de détails d'un signal $X(\omega)$ est illustré par la figure 3.4. Lorsque cette analyse est itérée N fois sur la composante d'approximation, on dispose des coefficients d'ondelettes aux échelles de 1 à N et du reste d'approximation à l'échelle N.

Analyse multirésolution. L'analyse multirésolution permet de trouver des ondelettes $\psi \in L_2(\mathbb{R})$ telles que la famille $\{\psi_{n,p}\}_{n,p\in\mathbb{Z}}$ de fonctions définies par :

$$\forall (n,p) \in \mathbb{Z}^2, \qquad \psi_{n,p}(t) = 2^{-p/2}\psi(2^{-p}t - n) \tag{3.8}$$

est une base orthonormée de $L_2(\mathbb{R})$. La transformée discrète en ondelettes de $f \in L_2(\mathbb{R})$ est donnée par l'ensemble des projections de f sur les fonctions de base :

$$\forall (n,p) \in \mathbb{Z}^2, \qquad c_{n,p}(f) = \langle f, \psi_{n,p} \rangle = \int f(t)\overline{\psi_{n,p}}(t)\mathrm{d}t \tag{3.9}$$

Si la base est orthonormée, f se reconstruit par la superposition des coefficients d'ondelettes :

$$f(t) = \sum_{n,p} c_{n,p}(f)\psi_{n,p}(t) \tag{3.10}$$

Une analyse multirésolution est une suite d'espaces vectoriels $(V_p)_{p\in\mathbb{Z}}$ vérifiant les propriétés suivantes :

(i) V_p est un sous espace fermé de $L_2(\mathbb{R})$
(ii) $V_p \subset V_{p-1}$
(iii) $\overline{\cup_{p\in\mathbb{Z}} V_p} = L_2(\mathbb{R})$ et $\cap_{p\in\mathbb{Z}} V_p = \{0\}$
(iv) $\exists \varphi \in V_0$ telle que $\{\varphi(t-n)\}_{n\in\mathbb{Z}}$ est une base orthonormée de V_0, cette fonction est la *fonction d'échelle* de l'analyse multirésolution.
(v) $\forall p \in \mathbb{Z},\ \ v(t) \in V_p \Leftrightarrow v(2t) \in V_{p-1}$

La projection orthogonale d'un signal f sur cette suite d'espaces emboîtés permet de construire une base orthonormée d'ondelettes. La relation (i) assure l'existence de la projection orthogonale de f sur chacun des V_p, et les relations (ii) et (iii) que la suite des projections converge vers f pour tout $f \in L_2(\mathbb{R})$. La démonstration dans [Mal89] suit plusieurs étapes. La première est d'utiliser les relations (ii), (iv) et (v) pour montrer que la famille $\{\varphi_{n,p}(t) = 2^{-p/2}\varphi(2^{-p}t - n)\}_{n\in\mathbb{Z}}$ est une base orthonormée de V_p. Ceci permet de définir l'approximation de f à la résolution p par :

$$Proj_{V_p}(f) = \sum_{n\in\mathbb{Z}} \langle f, \varphi_{n,p}\rangle \varphi_{n,p} \tag{3.11}$$

La deuxième étape consiste à montrer qu'il existe une fonction ψ telle que la famille $\{\psi_{n,p}\}_n$ est une base orthonormée du supplémentaire orthogonal de V_p dans V_{p-1}. On a alors la relation suivante :

$$Proj_{V_{p-1}}(f) = Proj_{V_p}(f) \ + \ \sum_{n\in\mathbb{Z}} \langle f, \psi_{n,p}\rangle \psi_{n,p} \tag{3.12}$$

Il s'en suit que la famille $\{\psi_{n,p}\}_{n,p\in\mathbb{Z}}$ est une base orthonormée de $L_2(\mathbb{R})$.

Analyse multirésolution et banc de filtres. Outre la construction de bases orthonormées d'ondelettes, l'analyse multirésolution permet le calcul des coefficients d'approximation et d'ondelettes par banc de filtres. Certaines propriétés de ce banc de filtres rendent compte de sa variance aux translations et seront discutées dans la section 3.2.1. Les fonctions d'échelle $\varphi_{n,p+1}$ se projettent sur la base $\{\varphi_{n,p}\}_n$ en :

$$\varphi_{n,p+1} = \sum_{k\in\mathbb{Z}} \langle \varphi_{n,p+1}, \varphi_{k,p}\rangle \varphi_{k,p} \tag{3.13}$$

Il est possible d'écrire le produit scalaire indépendamment de la résolution p :

$$\begin{aligned}\langle \varphi_{n,p+1}, \varphi_{k,p} \rangle &= \int 2^{(-p+1)/2} \varphi(2^{(-p+1)}t - n) 2^{-p/2} \varphi(2^{-p}t - k) \mathrm{d}t \\ &= \int 2^{-1/2} \varphi(\frac{t}{2}) \varphi(t - (k - 2n)) \mathrm{d}t \\ &= \langle \varphi_{0,1}, \varphi_{k-2n,0} \rangle\end{aligned}$$

L'approximation de f à la résolution $p+1$ s'écrit donc :

$$\langle f, \varphi_{n,p+1} \rangle = \sum_{k \in \mathbb{Z}} \langle f, \varphi_{k,p} \rangle h[k - 2n] \tag{3.14}$$

où $h[n] = \langle \varphi_{0,1}, \varphi_{n,0} \rangle = \int 2^{-1/2} \varphi(\frac{t}{2}) \varphi(t - n)) \mathrm{d}t$ est la réponse impulsionnelle d'un filtre passe-bas. Ainsi apparaît la possibilité d'obtenir récursivement les coefficients de l'approximation de f à l'échelle p (la projection de f sur V_p). En substituant le filtre h par son retourné temporel dans l'équation 3.14, l'approximation s_{p+1} de f à l'échelle p peut s'écrire comme la convolution entre le filtre h et le signal s_p, suivie d'une décimation de facteur deux :

$$s_{p+1}[n] = \langle f, \varphi_{n,p+1} \rangle = \sum_{k \in \mathbb{Z}} h[2n - k] s_p(k) \tag{3.15}$$

Le lien entre la fonction d'échelle continue φ et le filtre numérique h est donné en réécrivant l'équation 3.13 :

$$\varphi_{0,1}(t) = \frac{1}{\sqrt{2}} \varphi(\frac{t}{2}) = \sum_{k \in \mathbb{Z}} h[k] \varphi(t - k) \tag{3.16}$$

soit dans le domaine de Fourier :

$$\hat{\varphi}(2\omega) = H(\omega) \hat{\varphi}(\omega) = \hat{\varphi}(0) \prod_{n=0}^{+\infty} H(\frac{\omega}{2^n}) \tag{3.17}$$

où $H(\omega) = \sqrt{2} \sum_n h[n] e^{-in\omega}$. En combinant cette équation avec l'équation d'orthonormalité de la base $\{\varphi_{n,0}\}_n$ de V_0 :

$$\sum_{n \in \mathbb{Z}} |\hat{\varphi}(\omega + 2n\pi)|^2 = 1 \tag{3.18}$$

il vient l'équation de reconstruction exacte des filtres conjugués en quadrature :

$$|H(\omega)|^2 + |H(\omega + \pi)|^2 = 1 \tag{3.19}$$

Cette équation montre que le module du spectre de h est borné entre 0 et 1. Comme $\hat{\varphi}(0)$ est non nul (sinon $\hat{\varphi}$ est nul partout), $H(0) = 1$ (soit la condition de normalisation $\sum_n h[n] = \sqrt{2}$), et $H(\pi) = 0$. Le filtre h est donc bien un filtre passe-bas, permettant d'obtenir une version grossière du signal d'origine.

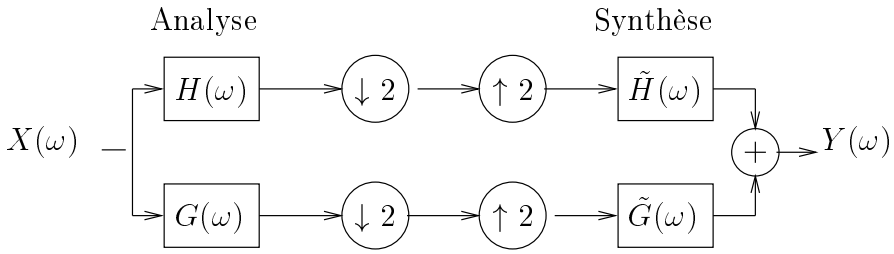

FIG. 3.4 – Analyse et synthèse par deux paires de filtres bi-orthogonaux.

De même que pour le filtre passe-bas, le filtre passe-haut $g[n]$ s'obtient en projetant $\psi_{0,1}$ sur V_0 :

$$g[n] = \langle \psi_{0,1}, \varphi_{n,0} \rangle = \int 2^{-1/2} \psi(\frac{t}{2}) \varphi(t-n)) dt \tag{3.20}$$

La condition d'orthonormalité de la famille $\{\psi_{n,0}\}_n$ conduit à la relation entre le filtre passe-bas et le filtre passe-haut :

$$g[n] = (-1)^n h[-n+1] \tag{3.21}$$

La relation entre l'ondelette continue ψ et le filtre numérique h est :

$$\psi(\frac{t}{2}) = \sqrt{2} \sum_{n \in \mathbb{Z}} (-1)^n h[-n+1] \varphi(t-n) \tag{3.22}$$

soit dans le domaine de Fourier :

$$\hat{\psi}(2\omega) = H^*(\omega + \pi) e^{-i\omega} \hat{\varphi}(\omega) = H^*(\omega + \pi) e^{-i\omega} \prod_{n=1}^{+\infty} H(\frac{\omega}{2^n}) \tag{3.23}$$

L'algorithme de décomposition en ondelettes par bancs de filtres permet de calculer récursivement les coefficients d'approximation $s_p[n]$ et d'ondelettes $c_p[n]$ par convolution et décimation :

$$\begin{aligned} s_p[n] &= \sum_k h[2n-k] s_{p-1}[k] \\ c_p[n] &= \sum_k g[2n-k] s_{p-1}[k] \end{aligned} \tag{3.24}$$

L'algorithme de reconstruction s'effectue par convolutions et insertions de zéros :

$$s_{p-1}[k] = \sum_n h[2n-k] s_p[n] \; + \; \sum_n g[2n-k] c_p[n] \tag{3.25}$$

La seule ondelette symétrique à support compact pouvant être générée par une telle analyse multirésolution est l'ondelette de Haar valant 1 sur [0,1/2[, -1 sur [1/2,1[, et 0 partout ailleurs. En revanche, une infinité d'ondelettes (avec un degré de régularité

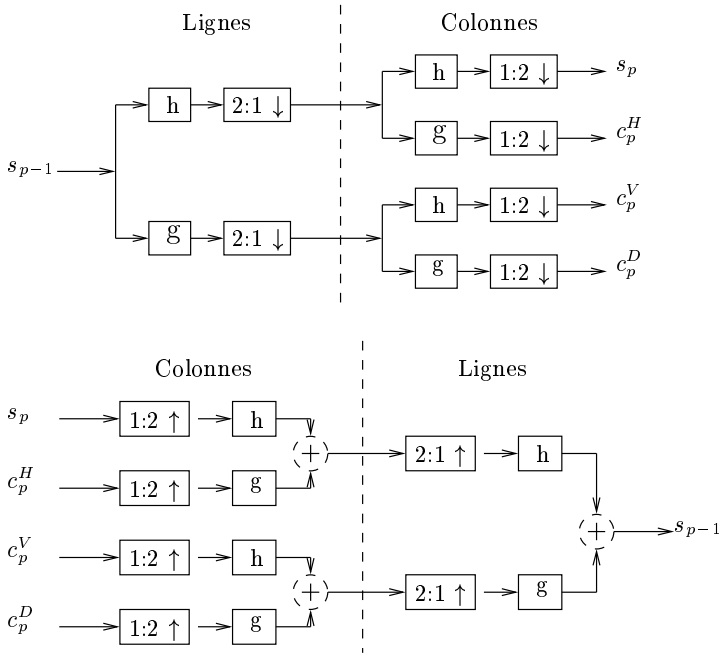

FIG. 3.5 – Un étage de la décomposition multirésolution bidimensionnelle (en haut), et un étage de la synthèse associée (en bas).

quelconque) peut être engendrée par analyse multirésolution en remplaçant la contrainte d'orthogonalité par une contrainte de bi-orthogonalité. Dans ce cas, la reconstruction du signal f donnée par l'équation 3.10 devient [Dau92] :

$$f = \sum_{n,p} \langle f, \psi_{n,p} \rangle \tilde{\psi}_{n,p} \tag{3.26}$$

où $\tilde{\psi}$ est l'ondelette duale de ψ, c'est-à-dire vérifie :

$$\langle \psi_{n,p}, \tilde{\psi}_{n',p'} \rangle = \delta_{n,n'} \delta_{p,p'} \tag{3.27}$$

L'analyse bi-orthogonale s'effectue par le banc de filtres de la figure 3.4 mettant en jeu deux paires de filtres bi-orthogonaux.

Extension au cas bidimensionnel. L'extension classique de la transformée discrète par ondelettes aux signaux bidimensionnels se fait par produit tensoriel. La fonction d'échelle est séparable dans le repère cartésien :

$$\varphi(x, y) = \varphi(x)\varphi(y) \tag{3.28}$$

L'approximation de $f(x, y)$ à la résolution 2^{-p} est alors donnée par :

$$s_p[n_x, n_y] = \langle f(x, y), \varphi_{n_x,p}(x)\varphi_{n_y,p}(y) \rangle \tag{3.29}$$

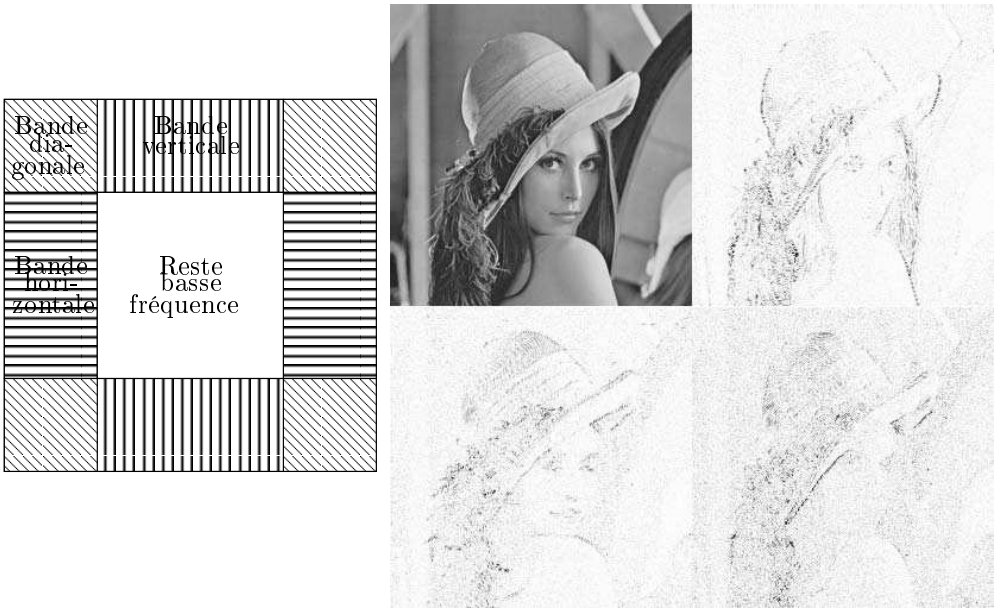

FIG. 3.6 – Figure de gauche : partition du plan espace-fréquence opérée par la transformée discrète en ondelettes séparables. Figure de droite : visualisation sur niveaux de gris normalisés de la valeur absolue des coefficients d'ondelettes séparables sur grille dyadique.

Comme dans le cas mono-dimensionnel, les coefficients d'ondelettes sont obtenus par projection de f sur le supplémentaire orthogonal de V_p dans V_{p-1} généré par translation et dilatation non plus d'une ondelette mère mais de trois ondelettes mères définies par :

$$\begin{aligned}\psi^H(x,y) &= \varphi(x)\psi(y)\\ \psi^V(x,y) &= \psi(x)\varphi(y)\\ \psi^D(x,y) &= \psi(x)\psi(y)\end{aligned} \tag{3.30}$$

Les coefficients d'ondelettes sont donc répartis dans trois sous-bandes à chaque niveau de résolution p :

$$\begin{aligned}c_p^H[n_x,n_y] &= \langle f(x,y), \varphi_{_{p,n_x}}(x)\psi_{p,n_y}(y)\rangle\\ c_p^V[n_x,n_y] &= \langle f(x,y), \psi_{p,n_x}(x)\varphi_{_{p,n_y}}(y)\rangle\\ c_p^D[n_x,n_y] &= \langle f(x,y), \psi_{p,n_x}(x)\psi_{p,n_y}(y)\rangle\end{aligned} \tag{3.31}$$

La figure 3.5 montre les bancs de filtres associés à une telle décomposition séparable par ondelettes, et la figure 3.6 permet de visualiser en niveaux de gris normalisés les coefficients d'ondelettes dans chacune des sous-bandes.

3.1.4 Quantification et codage des coefficients d'ondelettes

La figure 3.7 montre que la distribution des niveaux de gris des images naturelles est multimodale et de grande entropie, tandis que celle des coefficients d'ondelettes

est unimodale, centrée à l'origine et très étroite. Dans l'espace transformé, l'énergie est concentrée sur un faible nombre de coefficients, facilitant ainsi la compression. La distribution des coefficients d'ondelettes des images naturelles est modélisée par une loi gaussienne généralisée [Mal89] :

$$p(x) = Ke^{-\left(\frac{|x|}{\alpha}\right)^{\beta}} \tag{3.32}$$

où α contrôle la variance et β la vitesse de décroissance. Ces deux paramètres peuvent se calculer à partir des deux premiers moments de la distribution empirique. Cette loi peut être utilisée dans l'algorithme de Lloyd-Max visant à déterminer la quantification optimale au sens de la minimisation de la fonction débit-distorsion. En fait, la quantification uniforme est optimale pour les distributions laplaciennes, et donc proches de l'optimum pour les distributions de type gaussien généralisé. La quantification vectorielle permet d'obtenir de meilleurs résultats que la quantification scalaire, au prix de la génération d'un dictionnaire des vecteurs statistiquement les plus probables. Le problème de la constitution du dictionnaire permettant d'optimiser un compromis entre compression et description n'a jamais été traité, et est délicat. Dans une perspective de description, la contrainte de covariance aux similitudes des vecteurs du dictionnaire est difficile à respecter. Dans cette thèse, les coefficients seront donc uniformément quantifiés dans chacune des sous-bandes.

La figure 3.8 représente la distribution des coefficients d'ondelettes sachant une valeur spatialement adjacente dans la même bande ou sachant la valeur correspondant à la même position dans la bande de résolution supérieure. Une colonne donne en niveaux de gris la probabilité d'un coefficient d'ondelettes sachant que son voisin (en espace ou en échelle) est égale à la valeur donnée en abscisse. Les distributions de chaque colonne sont de moyenne nulle, conformément à la nature décorrélée des coefficients d'ondelettes. En revanche, la forme en noeud papillon des distributions conditionnelles montre que les coefficients adjacents ne sont pas indépendants. Cette forme spécifique est avantageusement exploitée dans les schémas de compression comme JPEG 2000 par les codeurs EZW [Sha93] ou similaires comme EBCOT. Leur principe consiste à stipuler que les fils d'un coefficient non signifiant (de valeur absolue inférieure à un seuil) sont non signifiants. Un tel codage est inadapté pour la description. Les transformations admissibles comme les similitudes ou l'ajout de bruit peuvent considérablement modifier l'arbre des coefficients signifiants.

3.2 Variance de la transformée discrète en ondelettes sur grille dyadique

La covariance aux similitudes de la transformée continue en ondelettes la rend pertinente pour le problème de description. Pour réduire sa redondance dans un but de compression, la section précédente a montré comment discrétiser la transformée pour ne conserver que le nombre minimal de coefficients nécessaire pour la reconstruction.

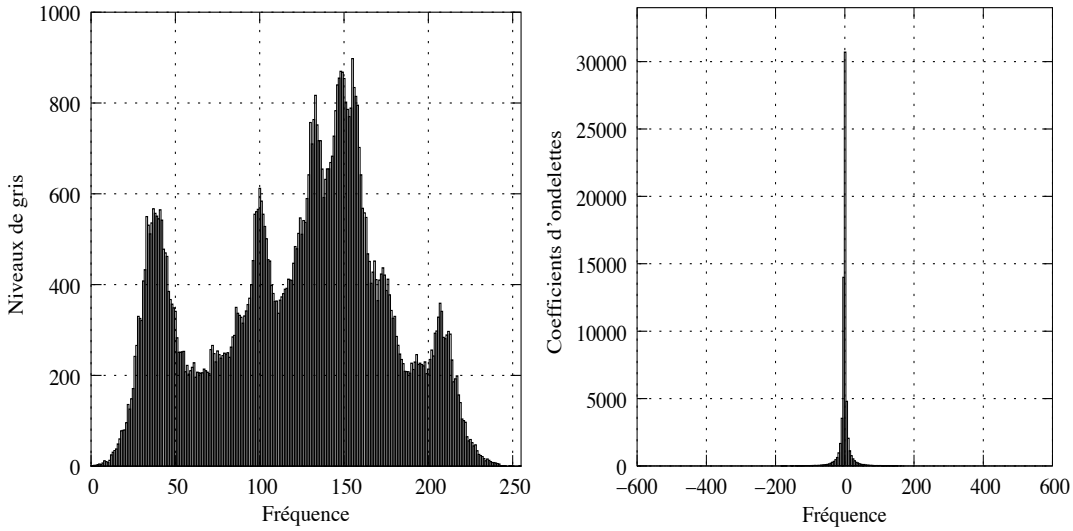

FIG. 3.7 – Distribution des niveaux de gris de l'image Lena (à gauche), et des coefficients d'ondelettes (à droite).

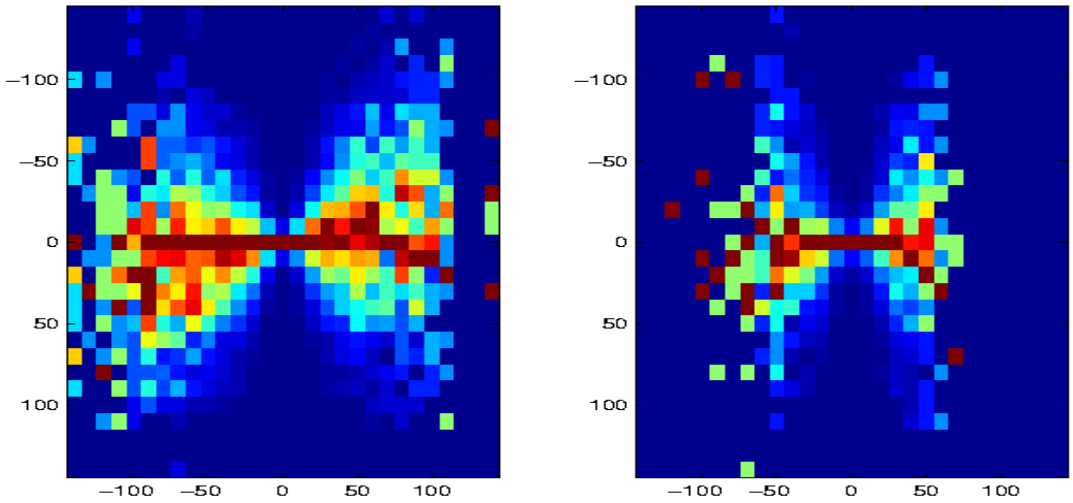

FIG. 3.8 – Distribution conditionnelle des coefficients d'ondelettes calculée à partir de l'image Lena. Chaque colonne donne en niveaux de couleurs la probabilité d'un coefficient d'ondelettes sachant que son voisin (en espace sur la figure de gauche, et en échelle sur la figure de droite) est égale à la valeur donnée en abscisse.

Cette section montre que cette discrétisation implique des recouvrements spectraux inter-bandes faisant perdre à la transformée ses propriétés de covariance.

3.2.1 Variance à la translation

La figure 3.9, reproduite à partir de [SAH92], montre l'effet d'une translation d'une unité de temps sur la distribution inter et intra bande des coefficients d'ondelettes d'un signal temporel constitué d'une ondelette de Daubechies dilatée d'un facteur deux. La translation a pour effet de disperser l'énergie dans toutes les bandes d'analyse. La détection de points d'intérêt dans le domaine ondelettes devient dès lors instable, comme le montrera la section 3.3.2. Cette instabilité provient du sous-échantillonnage apparaissant dans le banc de filtres de la figure 3.4. Pour le constater, le premier élément à considérer est la présence de hautes fréquences dans la bande de basse fréquence, et inversement. Si l'on note $X(z) = \sum_n x[n]z^{-n}$ la transformée en z du signal $x[n]$, le signal $X_{\downarrow 2}(z)$ obtenu en ne conservant que les indices pairs s'écrit :

$$\begin{aligned} X_{\downarrow 2}(z) &= \sum_n x[2n]z^{-n} \\ &= \frac{1}{2}\sum_n (1+(-1)^n)x[n]z^{-\frac{n}{2}} \end{aligned}$$

En posant $z = e^{j\omega}$, les transformées de Fourier $X_b(\omega)$ et $X_h(\omega)$ des signaux apparaissant respectivement dans les bandes de basse et de haute fréquence de la figure 3.4 s'écrivent :

$$X_b(\omega) = \frac{1}{2}(H(\frac{\omega}{2})X(\frac{\omega}{2}) + H(\frac{\omega}{2}+\pi)X(\frac{\omega}{2}+\pi)) \quad (3.33)$$

$$X_h(\omega) = \frac{1}{2}(G(\frac{\omega}{2})X(\frac{\omega}{2}) + G(\frac{\omega}{2}+\pi)X(\frac{\omega}{2}+\pi)) \quad (3.34)$$

La figure 3.10 représente le module de la transformée de Fourier à décroissance quadratique (comme le spectre des images naturelles) d'un signal d'entrée $X(\omega)$ périodique de période 2π , le signal filtré $G(\omega)X(\omega)$, la composante $G(\frac{\omega}{2})X(\frac{\omega}{2})$ du signal filtré et sous-échantillonné, et le signal de sortie $X_h(\omega)$. Il apparaît que, même pour un passe-haut idéal de fréquence nulle sur $[-\frac{\pi}{2} - \frac{\pi}{2}]$, le spectre du signal filtré et sous-échantillonné n'est pas nul sur cet intervalle de basses fréquences. Ceci provient du recouvrement spectral dû au sous-échantillonnage.

Avec les notations de la figure 3.4, le signal de sortie s'écrit :

$$\begin{aligned} Y(\omega) = & \frac{1}{2}X(\omega)(H(\omega)G(\omega) + \tilde{H}(\omega)\tilde{G}(\omega)) \\ + & \frac{1}{2}X(\omega+\pi)(H(\omega+\pi)G(\omega+\pi) + \tilde{H}(\omega+\pi)\tilde{G}(\omega+\pi)) \end{aligned} \quad (3.35)$$

Il s'en suit les deux conditions de reconstruction parfaite :

$$H(\omega)G(\omega) + \tilde{H}(\omega)\tilde{G}(\omega) = 2 \quad (3.36)$$

$$H(\omega+\pi)G(\omega+\pi) + \tilde{H}(\omega+\pi)\tilde{G}(\omega+\pi) = 0 \quad (3.37)$$

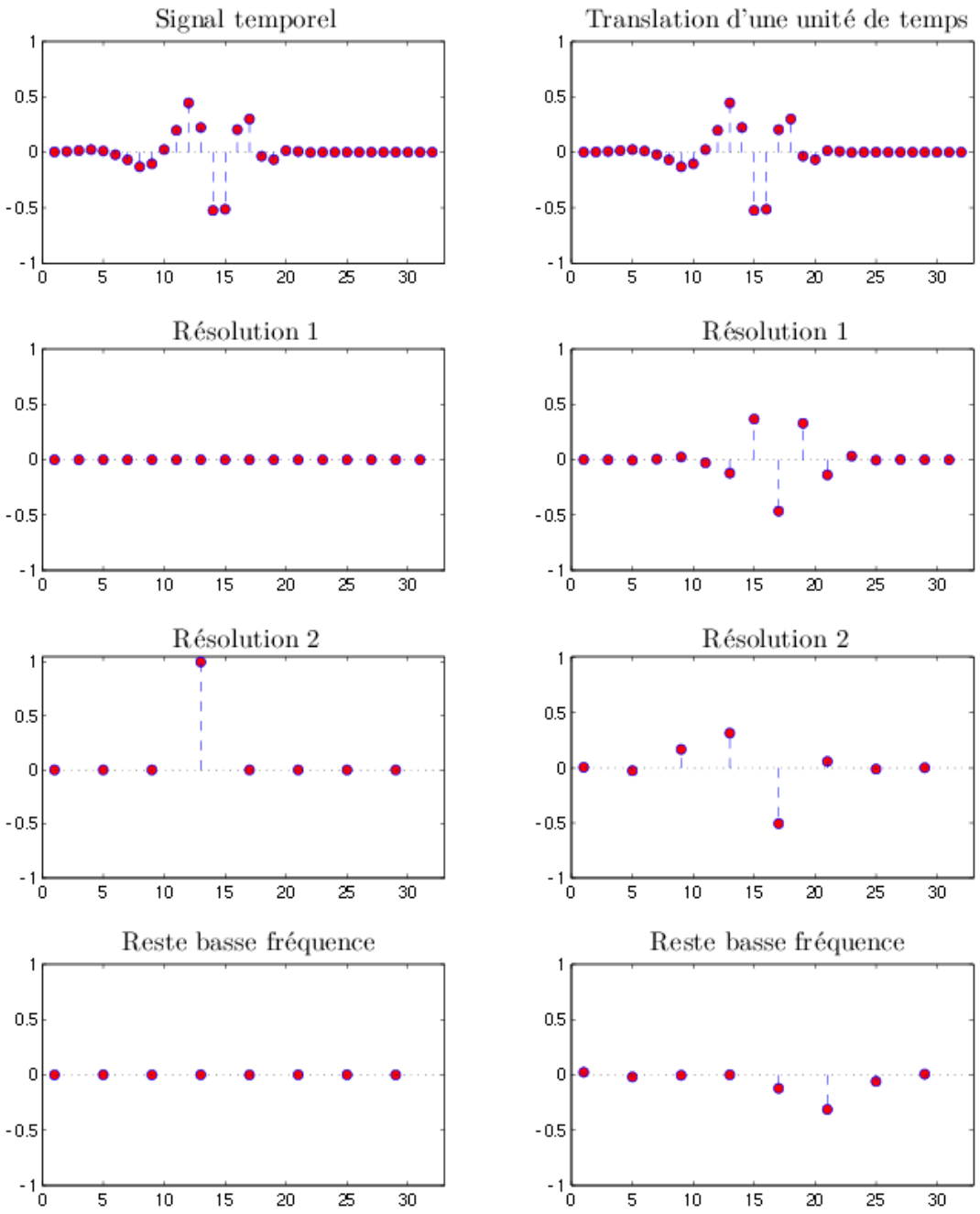

FIG. 3.9 – Première ligne : signal temporel (à gauche), et sa translatée d'une unité de temps (à droite). Leurs coefficients d'ondelettes se répartissent très différemment dans chacune des sous-bandes.

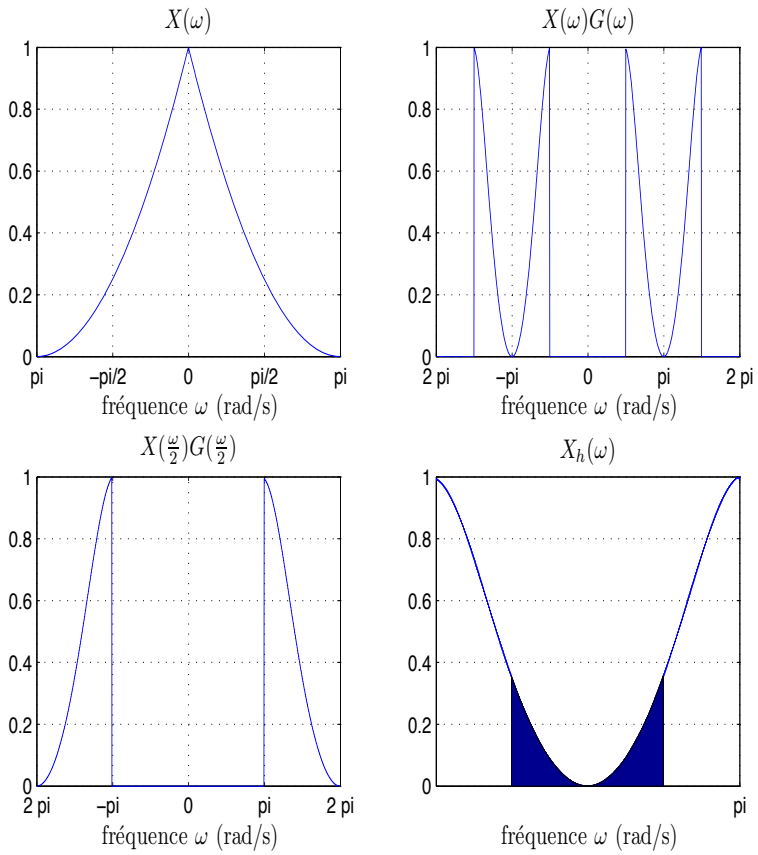

FIG. 3.10 – Signal d'entrée $X(\omega)$ (en haut à gauche), signal filtré $G(\omega)X(\omega)$ (en haut à droite), composante $G(\frac{\omega}{2})X(\frac{\omega}{2})$ du signal filtré et sous-échantillonné (en bas à gauche), et signal de sortie $X_h(\omega)$ comportant des basses fréquences en bleu (en bas à droite).

La condition 3.37 s'appelle la condition d'*anti-aliasing*. Elle assure que, *après* synthèse, les réplications spectrales causées par les sous-échantillonnages s'annulent. En revanche, dans toute bande d'analyse, la réplication engendre un recouvrement spectral. Ce recouvrement est responsable de la variance aux translations. Une façon de le constater est de considérer le signal reconstruit à partir d'une seule sous-bande. La figure 3.11 représente le banc de filtres en cascade utilisé pour le calcul de la transformée discrète en ondelettes. Le signal $X_k(\omega)$ est reconstruit uniquement à partir de la $k^{\text{è}}$ sous-bande $Y_k(\omega)$, les autres bandes étant mises à zéro. Un résultat classique en banc de filtres permet d'inverser l'ordre des convolutions et des décimations, ou plus précisément d'affirmer l'équivalence entre les deux systèmes de la figure 3.11. Si la bande conservée est la $k^{\text{è}}$ bande d'ondelettes, les filtres $A_k(\omega)$ et $S_k(\omega)$ sont donnés par :

$$\begin{aligned} A_k(\omega) &= H(\omega)H(2\omega)\dots H(2^{k-1}\omega)G(2^k\omega) \\ S_k(\omega) &= \tilde{H}(\omega)\tilde{H}(2\omega)\dots \tilde{H}(2^{k-1}\omega)\tilde{G}(2^k\omega), \end{aligned} \tag{3.38}$$

Si la bande conservée est le $n^{\text{è}}$ reste basse fréquence, les filtres équivalents se factorisent de la manière suivante :

$$\begin{aligned} A_n(\omega) &= H(\omega)H(2\omega)\dots\omega)H(2^n\omega) \\ S_n(\omega) &= \tilde{H}(\omega)\tilde{H}(2\omega)\dots \tilde{H}(2^n\omega) \end{aligned} \tag{3.39}$$

Le signal $X_k(\omega)$ s'écrit donc comme la somme d'une partie covariante dans le temps et d'une partie contenant les recouvrements spectraux :

$$X_k(\omega) = \frac{1}{2^k}A_k(\omega)S_k(\omega)X(\omega) \; + \; \frac{1}{2^k}\sum_{p=1}^{2^k-1} A_k(\omega+\frac{p\pi}{2^k})S_k(\omega)X(\omega+\frac{p\pi}{2^k}) \tag{3.40}$$

La $k^{\text{è}}$ sous-bande est donc variante aux translations, puisque la réponse du signal modulé en fréquence n'est pas le modulé de la réponse. La covariance aux translations dans chaque bande est équivalente à :

$$\forall 1 \leq k \leq n, \forall 1 \leq p \leq 2^k - 1, \quad A_k(\omega + \frac{p\pi}{2^{k-1}})S_k(\omega) = 0 \tag{3.41}$$

La figure 3.12 permet de visualiser le recouvrement entre les bandes $\{A_2(\omega + \frac{p\pi}{2^2})\}_{0\leq p\leq 7}$ et la bande $S_2(\omega)$ lorsque les filtres H et G des équations 3.38 et 3.39 sont ceux associés à l'ondelette de Daubechies de longueur 8. Pour le filtre passe-bas, la relation 3.41 est mise en défaut « seulement »pour $p = \pm 1$. Le recouvrement spectral est beaucoup plus important dans les bandes d'ondelettes, et a lieu pour $p = \pm 1$, et surtout pour $p = \pm 2$ Les filtres H et G étant réels et symétriques, les spectres de $A_k(\omega)$ et $S_k(\omega)$ sont symétriques par rapport à l'origine, impliquant un fort recouvrement. Une solution consiste à utiliser des filtres à spectre analytique, donc complexes [Kin98]. Il est important de souligner que la nature des recouvrements observés par l'ondelette de Daubechies est la même pour tout type d'ondelette. En effet, d'après les relations 3.19 et 3.21, les spectres

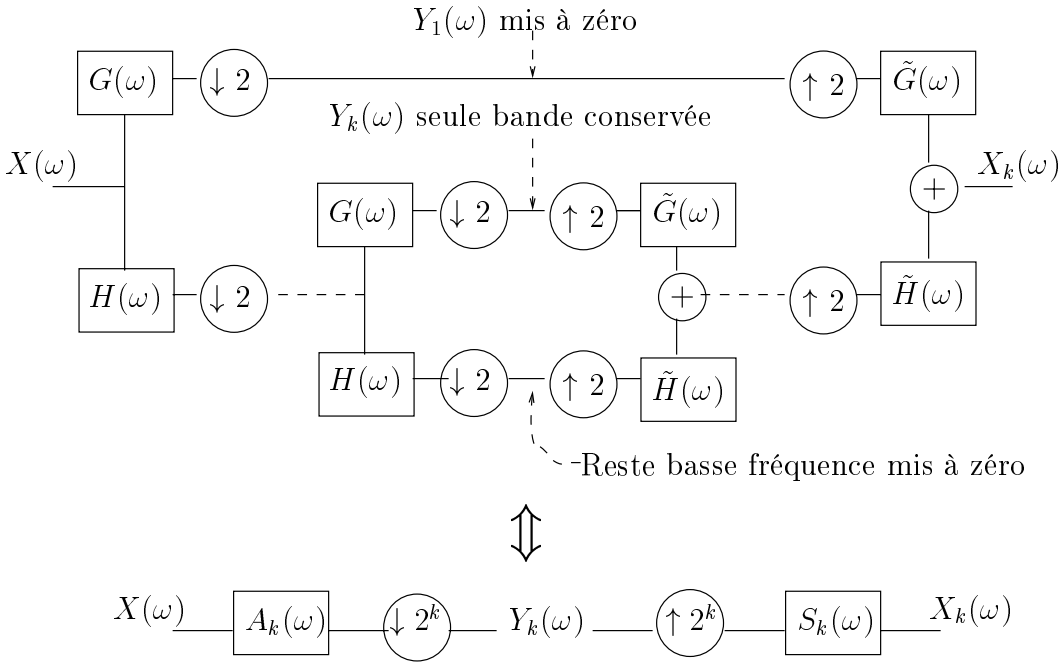

FIG. 3.11 – Banc de filtres en cascades. La reconstruction dans le schéma du haut s'effectue en ne conservant qu'une seule bande d'analyse, et en mettant à zéro les autres bandes. Ce schéma de reconstruction est équivalent à celui du bas, où les ordres de convolution et de décimation ont été inversés.

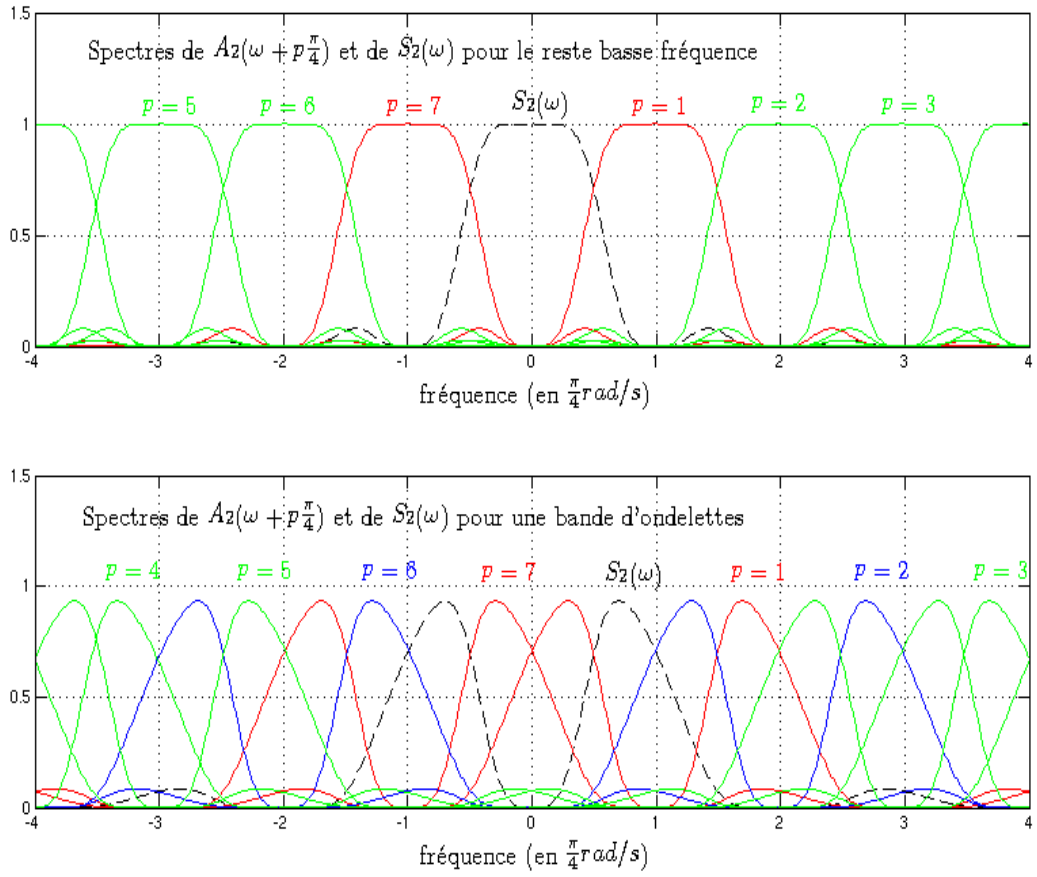

FIG. 3.12 – Recouvrements spectraux responsables de la variance aux translations. Le sous-échantillonnage induit deux types de recouvrement : un premier (en haut) entre les filtres passe-bas, un second (en bas) entre les filtres passe-haut.

de H et G sont d'énergie complémentaire, ce qui contraint fortement leur forme. En revanche, l'ampleur des recouvrements est variable selon le choix de l'ondelette. Pour s'en apercevoir, il suffit en effet de considérer l'ondelette de Shannon (la fonction *sinc* modulée en fréquence) pour laquelle les recouvrements sont nuls. Cette ondelette pose toutefois problème : elle n'est pas à support compact, elle ne peut pas être générée par des filtres à réponse impulsionnelle finie. Dans la section 3.3.2 sont testées plusieurs ondelettes dans un but de détection robuste de points d'intérêt.

3.2.2 Variance à la rotation

La section précédente a montré que les coefficients d'ondelettes sont très dépendants de la position relative de l'image par rapport à la grille dyadique sur laquelle a lieu la décomposition. Une rotation modifie en tout point et non uniformément cette position relative. Ceci constitue une première source de variance aux rotations. Une deuxième provient du produit tensoriel créant trois bandes dont l'énergie est maximale pour des contours orientés dans la direction horizontale, verticale ou l'une des deux directions diagonales. Pour évaluer la sensibilité de chacune de ces bandes à la direction des contours, on considère l'image synthétique I dont chacune des lignes est une fonction porte, et la famille d'images I_θ constituée des images I tournées d'un angle θ. À gauche de la figure 3.13 est représentée l'évolution du maximum d'énergie dans chacune des trois bandes en fonction de l'angle θ, et à droite l'énergie dans la bande verticale pour $\theta = 15$ degrés. Il apparaît que, dans le domaine ondelettes, l'énergie est fortement oscillatoire le long du contour, alors que, dans le domaine spatial, le contour est invariant par translation dans cette direction. Ceci pose un problème de description : la recherche d'invariants, scalaires ou calculés à partir de distributions, est difficile, sinon impossible. Un point plus rassurant est que la répartition de l'énergie dans chacune des bandes est à peu près identique pour l'ensemble des modèles de contours représentés sur la figure 3.14. Cette invariance aux modèles de contours permet d'espérer l'existence d'une énergie, définie comme somme pondérée des trois bandes $\{b_i\}_{1\leq i\leq 3}$, qui soit peu variante aux rotations, même pour les images naturelles.

Pour le vérifier, on cherche les poids $\{\alpha_i\}_{1\leq i\leq 3}$ permettant à la somme pondérée de minimiser l'erreur quadratique avec la fonction constante μ :

$$\{\alpha_1, \alpha_2, \alpha_3\} = \arg \min_{\{\alpha_1,\alpha_2,\alpha_3\}} \sum_{k=1}^{n} \left(\sum_{i=1}^{3} \alpha_i b_i(\theta_k) - \mu \right)^2 \tag{3.42}$$

où $\{\theta_k\}_{1\leq k\leq n}$ est l'ensemble des angles pour lesquels les énergies $\{b_i(\theta_k)\}_{1\leq i\leq 3, 1\leq k\leq n}$ ont été calculées à partir de l'ensemble des modèles de contours illustrés sur la figure 3.14. La fonction constante μ est la moyenne suivante :

$$\mu = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{3} \alpha_i b_i(\theta_k) \tag{3.43}$$

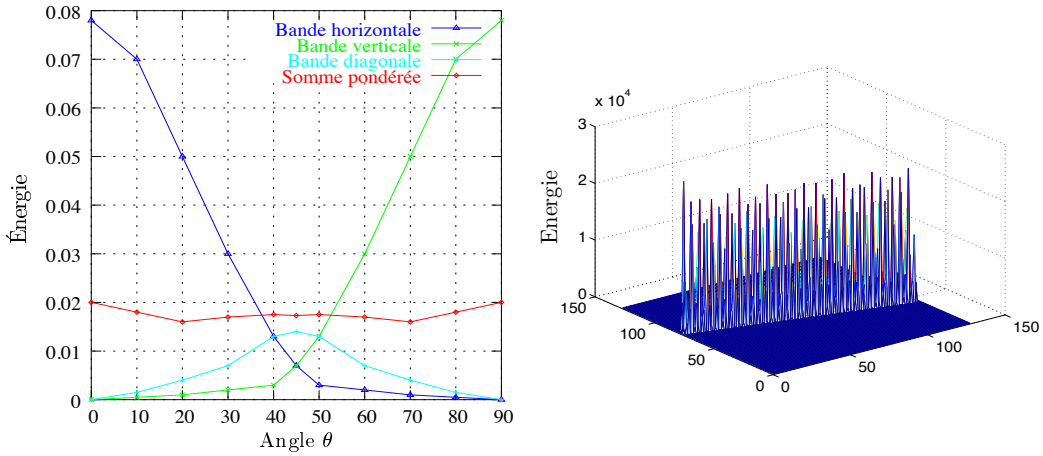

FIG. 3.13 – À gauche : répartition de l'énergie dans chaque bande d'ondelettes en fonction de l'orientation du contour. À droite : énergie dans la bande verticale pour un contour modélisé par une fonction porte faisant un angle de 15 degrés avec l'horizontale.

En dérivant l'erreur quadratique par rapport aux α_i cherchés, on obtient le système d'équations :

$$\forall 1 \leq j \leq 3,\ \sum_{k=1}^{n} b_j(\theta_k) \sum_{i=1}^{3} \alpha_i b_i(\theta_k) - \mu \sum_{k=1}^{n} b_j(\theta_k) = 0 \tag{3.44}$$

En posant E la matrice 3×3 dont les coefficients sont $e_{jl} = \sum_k b_j(\theta_k) b_l(\theta_k)$, et $\mathbf{S}$ le vecteur colonne de coefficients $s_j = \sum_{k=1}^{n} b_j(\theta_k)$, le vecteur $\alpha^T = (\alpha_1, \alpha_2, \alpha_3)$ cherché est solution de :

$$\mathbf{E}.\alpha = \mu \mathbf{S} \tag{3.45}$$

conduisant à $\alpha_1 = \alpha_2 = 0.18$ et $\alpha_3 = 0.64$, lorsque les $b_j(\theta_k)$ sont moyennés à partir des modèles de contours de la figure 3.14, et lorsque l'ondelette d'analyse est l'ondelette de Daubechies de longueur 8. Sur la figure 3.13 est représentée la somme ainsi pondérée en fonction de l'angle θ d'une fonction porte avec l'horizontale. Cette somme pondérée est peu variante aux rotations pour tous les modèles de contours illustrés sur la figure 3.14. La section suivante montrera que cette énergie est en fait très variante aux rotations pour les images naturelles.

3.3 Description dans le domaine compressé JPEG 2000

La transformée en ondelettes a permis des gains de codage importants, conduisant au nouveau standard de compression JPEG 2000. Il consiste en une transformée à échantillonnage critique sur grille dyadique par ondelettes séparables, d'une quantification scalaire uniforme, suivie d'un codage par plans de bits permettant la construction d'un

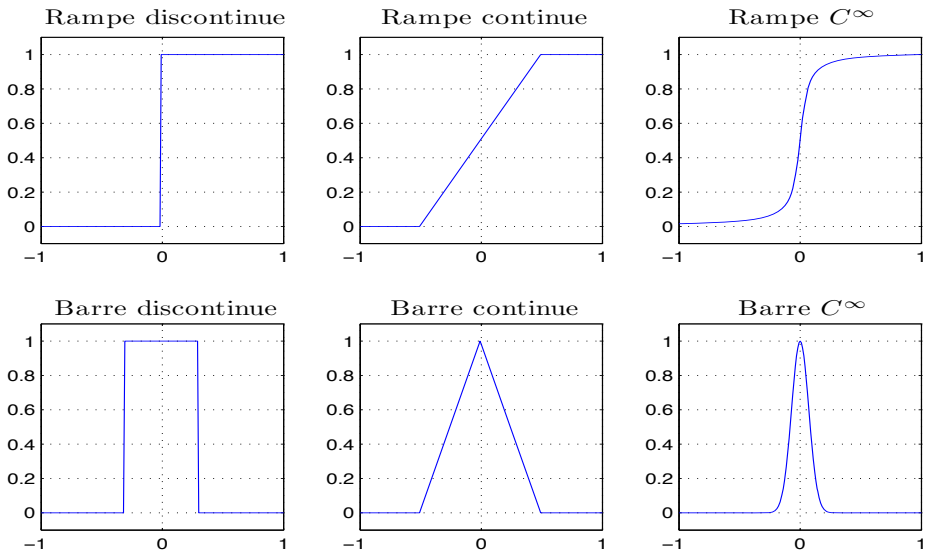

FIG. 3.14 – Modèles de contours utilisés pour l'estimation de la variance aux rotations.

arbre de zéros réduisant considérablement la dépendance inter-échelle [Sha93]. Concernant le problème de la description dans le domaine compressé JPEG 2000, la littérature n'a traité que la description globale et l'extraction robuste de points d'intérêt. Les principaux travaux sont présentés dans cette section. Quelques améliorations sont apportées au détecteur de points existant.

3.3.1 Description globale

Le sous-échantillonnage, responsable de la forte variance des coefficients transformés aux translations et aux rotations, incite à une description globale et non pas locale. La description globale consiste en la description de chacune des sous-bandes d'analyse. La section précédente a montré qu'une translation d'un seul pixel redistribue l'énergie dans chacune des bandes, ce qui limite la robustesse possible de la description, même globale. L'intérêt de la description dans le domaine compressé étant important, des travaux ont tout de même cherché à extraire des caractéristiques globales robustes et discriminantes. La motivation de ces travaux réside dans la possibilité d'effecteur des traitements visuels en économisant une partie du temps de décompression. Il est important de rappeler que le temps de décodage de l'arbre de zéros ne peut pas être épargné, seul le temps de calcul de la transformée inverse peut l'être. Certains travaux [Lin97, VJ02] ont tout de même tenté d'évaluer la similarité entre images en comparant directement leurs arbres de zéros. Ces techniques sont très rapides, mais aux résultats peu fiables et ne s'appliquent qu'à des images de même taille (sans quoi les arbres ne sont pas comparables). La majorité des travaux s'effectue dans le domaine ondelettes, après l'opération de décodage. Les signatures les plus simples ne décrivent que les statistiques de premier ordre des

coefficients d'ondelettes. Les plus courantes sont la moyenne et la variance de l'énergie dans chaque bande [KC92], et les paramètres α et β de la gaussienne généralisée de la relation 3.32 permettant de modéliser la distribution des coefficients dans chaque bande [dWSLD99]. La recherche dans la base de l'image la plus similaire s'effectue fréquemment par minimisation de l'erreur quadratique entre les signatures extraites. Dans [DV02] est proposée une recherche par maximum de vraissemblance, dont le coût de calcul est faible dans le cas où les signatures sont les paramètres α et β modélisant la distribution des coefficients d'ondelettes. Des signatures plus élaborées, prenant en compte les statistiques de second ordre ont été proposées dans [dWSD99] et apportent un gain en classification de texture.

3.3.2 Description locale

Ondelettes et détecteur de Canny. La forte variance aux translations et rotations de la transformée discrète en ondelettes a considérablement limité les travaux portant sur la description locale. Les seules recherches portent sur l'extraction robuste de points d'intérêt. Le point de départ de ces travaux est le lien effectué dans [MH92] entre la transformée continue en ondelettes et le détecteur de Canny. Si l'on choisit une gaussienne de variance unité comme fonction d'échelle $\phi(x, y)$, les dérivées partielles

$$\psi_1(x, y) = \frac{\partial \phi(x, y)}{\partial x} \text{ et } \psi_2(x, y) = \frac{\partial \phi(x, y)}{\partial y}$$

sont des ondelettes, car d'intégrales nulles, et sont de régularité infinie. La symétrie de la fonction d'échelle permet d'écrire les transformées en ondelettes de l'image I par ψ_1 et ψ_2 comme des convolutions :

$$\begin{pmatrix} T_1[I](x, y, s) \\ T_2[I](x, y, s) \end{pmatrix} = s \begin{pmatrix} \frac{\partial}{\partial x}(I \star \phi_s)(x, y) \\ \frac{\partial}{\partial y}(I \star \phi_s)(x, y) \end{pmatrix} = s \nabla (I \star \phi_s)(x, y) \tag{3.46}$$

où ϕ_s est la gaussienne d'écart-type s. Les coefficients d'ondelettes permettent donc d'extraire les contours de Canny correspondant à l'ensemble des points où le module du gradient est localement maximum dans la direction du gradient.

Ondelettes et points d'intérêt. La détection de points d'intérêt dans le domaine dyadique n'a été étudiée que par Loupias dans [LS99]. L'algorithme est simple et rapide. Les points sont initialisés à partir des coefficients de la plus grossière résolution. Chacun de ces coefficients est le père des coefficients de la bande juste plus fine qui correspondent à la même localisation spatiale. Le nombre de fils dépend de la largeur des filtres utilisés. En parcourant les résolutions de plus en plus fines, chaque point est remplacé par le fils d'énergie maximale, si bien que leur localisation s'affine progressivement. La figure 3.15 montre que la répétabilité de cet extracteur est très en-deçà de celle de Harris-Laplace présentée en 2.2, particulièrement pour les faibles erreurs de localisations.

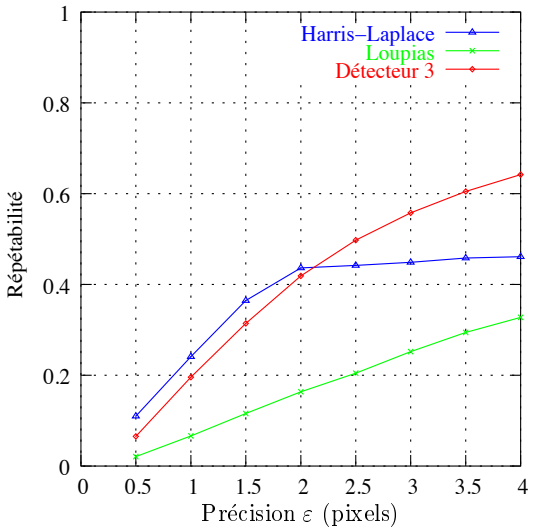

FIG. 3.15 – Répétabilités comparées du détecteur de Harris-Laplace, de Loupias, et d'un nouveau détecteur, pour la détection de copies. Les copies sont des images ayant subi une dilatation de facteur 1.6, une rotation de 30 degrés, une compression JPEG de facteur 10, et un crop de 30%.

Dans la suite de la thèse seront proposés des détecteurs de points dans des espaces transformés redondants. Il sera donc intéressant de les comparer avec leur équivalent dans le domaine ondelettes séparables à échantillonnage critique.

- Lorsque la représentation est sous-échantillonnée, les détecteurs reposent tous sur l'extraction de maxima locaux (en espace seulement) d'énergie dans le domaine transformé. L'extraction de maxima locaux en espace et en échelle n'a pas de sens immédiat pour des représentations sous-échantillonnées ; les tentatives qui ont été faites n'ont pas permis d'extraire des points robustes. Les maxima locaux sont extraits à chaque échelle et sont repositionnés, ou recalés, sur la grille de pleine résolution.
- L'énergie dont les points d'intérêt sont des maxima locaux peut être définie de différentes façons. Elle peut être le carré des coefficients dans chaque bande, conduisant à extraire les points indépendamment dans les bandes horizontale, verticale et diagonale. Elle peut également être une somme, équipondérée ou pondérée comme dans la section 3.2.2, des énergies de chaque bande.
- Différentes méthodes sont envisageables pour positionner les maxima locaux, extraits à une certaine échelle, sur la grille de pleine résolution sur laquelle est définie l'image. Deux méthodes de repositionnement ont été retenues : la première consiste à simplement transformer le maximum local $\mathbf{x}$ extrait à l'échelle p en $2^p\mathbf{x}$ sur la grille de pleine résolution (recalage de type 1) ; la seconde s'inspire de [LS99] et consiste à récursivement substituer les points extraits par leur fils d'énergie maximale, jusqu'à obtenir des points définis sur la grille de résolution

	Shannon	9-7	fmq 9	Daubechies 4	Haar
Détecteur 1	**0.09**	0.15	0.17	0.10	0.26
Détecteur 2	0.08	0.16	0.14	0.09	0.24
Détecteur 3	0.08	**0.17**	**0.19**	**0.11**	**0.31**
Détecteur 4	0.07	0.12	0.13	0.08	0.27
Détecteur 5	0.07	0.11	0.13	0.08	0.24
Détecteur 6	0.08	0.14	0.15	**0.11**	0.29

FIG. 3.16 – Répétabilité à 1.5 pixels pour les six détecteurs et pour différentes ondelettes.

la plus fine (c'est-à-dire de résolution moitié), qui sont ensuite extrapolés sur la grille de pleine résolution (recalage de type 2).

Les trois définitions possibles de l'énergie à chaque échelle, et les deux différentes méthodes de recalage conduisent aux six détecteurs suivants :

- Détecteur 1 : une énergie par échelle, définie comme somme équipondérée et recalage de type 1;
- Détecteur 2 : une énergie par échelle, définie comme somme pondérée et recalage de type 1;
- Détecteur 3 : trois énergies par échelle, et recalage de type 1;
- Détecteur 4 : une énergie par échelle, définie comme somme équipondérée et recalage de type 2;
- Détecteur 5 : une énergie par échelle, définie comme somme pondérée et recalage de type 2;
- Détecteur 6 : trois énergies par échelle, et recalage de type 2.

Dans le cas du recalage de type 1, l'échelle du point est simplement l'échelle où le point a été détecté comme maximum local. Dans le cas du recalage de type 2, l'échelle du point est l'échelle où l'énergie est maximale en parcourant les fils jusqu'à la grille de résolution moitié.

Répétabilité des extracteurs de points. Le tableau 3.3.2 permet de comparer la répétabilité de ces six extracteurs, pour différentes ondelettes. Les ondelettes testées sont l'ondelette de Shannon, qui d'après 3.2.1 est l'ondelette minimisant le recouvrement spectral dû au sous-échantillonnage, l'ondelette 9-7 utilisé dans le standard JPEG 2000, l'ondelette miroir en quadrature avec la fonction d'échelle de longueur 9, l'ondelette de Daubechies de longueur 8, et l'ondelette de Haar. Certains résultats infirment les prédictions.

- L'ondelette de Shannon, attendue comme celle conduisant aux points les plus robustes, est au contraire parmi les plus mauvaises. La meilleure ondelette est la simple ondelette de Haar. Cela peut s'expliquer par le fait que, pour minimiser les recouvrements spectraux inter-bandes, l'ondelette de Shannon est fortement oscillante et à décroissance très lente, donc inadaptée à la détection de points saillants. La bonne robustesse des points extraits par l'ondelette de Haar montre

que le choix de l'ondelette doit être plus guidé par sa forme propice à la détection de saillance, que par la minimisation des recouvrements spectraux. Le prochain chapitre montrera néanmoins que la minimisation des recouvrements est importante dans le choix de la discrétisation du plan espace-fréquence associée à une analyse espace-échelle donnée.

- La symétrie des filtres est importante pour la détection. En effet, les filtres à phase non linéaires ne permettent pas la localisation des primitives saillantes. Ainsi, l'ondelette de Daubechies conduit à de mauvais résultats.
- Les résultats présentés dans le tableau 3.3.2 montrent que les trois types d'énergie sont sensiblement équivalentes. Une différence apparaît en revanche sur le type de recalage des points sur la grille de pleine résolution. Le recalage par simple extrapolation est meilleur que le recalage sélection hiérarchique des meilleurs fils.

Le meilleur détecteur est celui utilisant l'ondelette de Haar, trois énergies par échelle, et un recalage par simple extrapolation sur la grille de pleine résolution. La figure 3.15 permet de comparer sa répétabilité avec celle de Harris-Laplace et celle de Loupias. Ce détecteur est plus répétable mais moins précis que le détecteur de Harris-Laplace. La précision à laquelle ce détecteur devient meilleur que le détecteur de Harris-Laplace est de 2 pixels. Il apparaît donc qu'une extraction robuste de points est possible dans les représentations en ondelettes à échantillonnage critique, malgré leur forte variance aux translations. Cela est rendu possible par le fait que les coefficients les plus énergétiques sont beaucoup plus stables qu'un coefficient quelconque, comme le montre la comparaison des répétabilités entre ce détecteur et celui de Loupias.

Description des points d'intérêt. Le problème de variance aux translations est plus patent dans la phase de description des voisinages des points extraits. Pour transposer la description SIFT dans les représentations en ondelettes séparables, la relation 3.46 incite à définir l'orientation $\theta(\mathbf{x}, p)$ au point $(\mathbf{x}, p)$ de l'espace-échelle par :

$$\theta(\mathbf{x}, p) = \text{atan2}(c_p^V(2^{-p}\mathbf{x}), c_p^H(2^{-p}\mathbf{x})) \tag{3.47}$$

où atan2 est la fonction définie par la relation 2.3 (page 58), c_p^H et c_p^V sont respectivement les bandes horizontales et verticales à la résolution p, et 2^{-p} le facteur permettant de repositionner correctement le point $\mathbf{x}$ en tenant compte du sous-échantillonnage (recalage de type 1). Sur une image tournée de 30 degrés, le pourcentage d'orientations correctement détectées à 15 degrés près est très faible :

$$\frac{r_{p,s,\theta}(1.5, 1.6, 15)}{r_{p,s}(1.5, 1.6)} = 0.12$$

où $r_{p,s}(1.5, 1.6)$ est le pourcentage de points extraits dont l'erreur en localisation est inférieure à 1.5 pixels et l'erreur en échelle inférieure à une octave ; et $r_{p,s,\theta}(1.5, 1.6, 15)$ le pourcentage de points qui, en plus, ont une erreur en orientation inférieure à 15 degrés. Ce taux de 0.12 n'est que trois fois supérieur à $\frac{15}{360}$ qui correspond au taux d'une estimation aléatoire des orientations. On verra dans la section 4.1 que la méthode d'estimation

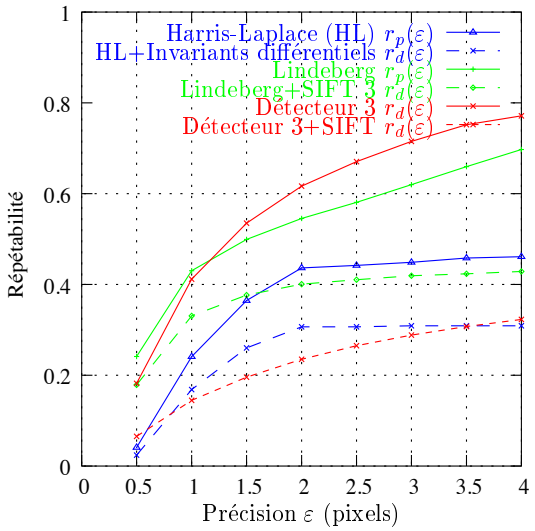

FIG. 3.17 – Répétabilités comparées du détecteur de Harris-Laplace, de Loupias, et du détecteur ondelettes dans les bande de basse fréquence. Les copies sont des images ayant subi une dilatation de facteur 1.6, une rotation de 30 degrés, une compression JPEG de facteur 10, et un crop de 30%.

en orientation définie par la relation 3.47 devient beaucoup plus robuste lorsqu'elle est effectuée à partir de représentations en ondelettes non sous-échantillonnées, où l'on a :

$$\frac{r_{p,s,\theta}(1.5, 1.6, 15)}{r_{p,s}(1.5, 1.6)} = 0.71$$

Cette instabilité d'estimation empêche la transposition des techniques de description reposant sur l'orientation comme la description SIFT.

Description avec reconstruction. Pour détecter des orientations de façon robuste, une solution consiste à reconstruire progressivement les restes de basse fréquence. Se donnant B_p la bande de basse fréquence à la résolution p, l'orientation $\theta(\mathbf{x}, p)$ au point $\mathbf{x}$ et à la résolution p peut se définir par:

$$\theta(\mathbf{x}, p) = \text{atan2}\left(\frac{\partial B_p(2^{-p}\mathbf{x})}{\partial x}, \frac{\partial B_p(2^{-p}\mathbf{x})}{\partial y}\right) \tag{3.48}$$

L'énergie servant de pondération de la contribution des orientations au calcul du descripteur SIFT est définie par:

$$E(\mathbf{x}, p) = \sqrt{\left(\frac{\partial B_p(2^{-p}\mathbf{x})}{\partial x}\right)^2 + \left(\frac{\partial B_p(2^{-p}\mathbf{x})}{\partial y}\right)^2} \tag{3.49}$$

L'orientation et l'énergie définies en tout point permettent de calculer le descripteur SIFT tel que présenté dans la section 2.4 (page 62). La figure 3.17 montre que la répétabilité des points extraits est très bonne, meilleure que celle de Harris-Laplace d'environ

50%, et de l'ordre de celle du détecteur de Lindeberg dédié aux blobs. Pour Harris-Laplace et Lindeberg, l'espace-échelle gaussien a été discrétisé géométriquement avec une raison de 1.1. Il est donc possible d'extraire des points tout aussi robustes avec une raison de 2 (au prix de la reconstruction des bandes de basse fréquence). En revanche, les descripteurs SIFT calculés dans les bandes de basse fréquence sont peu robustes. Ceci provient de la variance de la représentation, rendant difficile l'extraction d'échelles et d'orientations robustes. Cette technique d'extraction de caractéristiques dans les bandes reconstruites sera reprise dans le chapitre suivant, avec des représentations redondantes, moins variantes aux transformations géométriques admissibles.

3.4 Conclusion

Les représentations continues en ondelettes, par leur covariance aux similitudes, sont des candidates naturelles pour la description. La discrétisation classique de l'espace ondelette, conduisant aux représentations à échantillonnage critique telles que celles utilisées dans le standard JPEG 2000, ne permettent pas d'élaborer des schémas efficaces de description. Il est possible d'obtenir de bonnes performances pour l'extraction robuste de points en utilisant des ondelettes irrégulières comme l'ondelette de Haar. La robustesse des points extraits est même excellente si l'on accepte de reconstruire les bandes de basse fréquence. Il n'est en revanche pas possible d'estimer, dans le domaine transformé, des orientations robustes à partir des bandes échantillonnées construites par produit tensoriel. Le prochain chapitre étudie des représentations redondantes où une telle estimation est possible, permettant la construction de descripteurs efficaces inspirés des descripteurs SIFT.

Chapitre 4

Description à partir de représentations redondantes en ondelettes

Le troisième chapitre a étudié les représentations discrètes en ondelettes séparables définies sur grille dyadique à échantillonnage critique. Ces représentations, compactes et parcimonieuses, ont permis des gains considérables en compression. Les familles d'ondelettes, construites par translation et dilatation, constituent un outil naturel pour analyser l'ensemble des images d'une même scène prises dans des conditions variables de pose. Leur application directe au problème de compression et description simultanées est pourtant décevante. D'une part, le sous-échantillonnage critique, qui permet de ne conserver que le nombre minimal de coefficients pour la reconstruction, implique un recouvrement spectral inter-bandes responsable d'une forte variance aux translations. D'autre part, l'analyse anisotrope par produit tensoriel est sous-optimale en compression, et inadapté à de nombreux traitements visuels. Il s'en suit les conclusions tirées dans le chapitre précédent : s'il est possible d'obtenir des points d'intérêt robustes, il semble en revanche difficile d'estimer des orientations robustes, rendant impossible la transposition du descripteur SIFT dans le domaine ondelettes.

Des représentations redondantes ont été proposées pour remédier à ces problèmes. La représentation en ondelettes isotropes et non sous-échantillonnée [HKMMT89] est intéressante en description. Pour des applications requérant une redondance moindre, sa version sous-échantillonnée, la pyramide laplacienne, peut lui être substituée. L'absence d'information directionnelle est toutefois une limitation importante aux représentations isotropes. De nombreuses représentations directionnelles ont été construites dans la dernière décennie, comme les représentations séparables en ondelettes complexes [Kin98] ou leur généralisation par phaselets [Gop03] ; les représentations non séparables en curvelets [CD99], ridgelets [DV00], contourlets [Do01], directionlets [VBLVD], et les représentations orientables [FA91]. Ces représentations ont été appliquées à la compression et au débruitage, quelquefois à la description globale, jamais à la description locale.

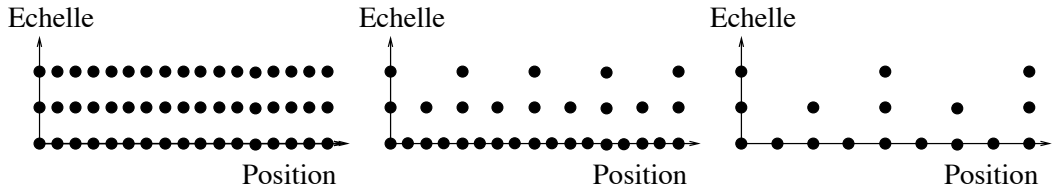

FIG. 4.1 – Partitions courantes du plan espace-fréquence : représentation dyadique non décimée (à gauche) ; représentation dyadique sans sous-échantillonnage à la première échelle (au milieu) ; représentation dyadique à échantillonnage critique (à droite).

La redondance de ces représentations est souvent liée à la partition du plan espace-fréquence qu'elles réalisent. La figure 4 donnent les représentations les plus communes. Pour chacune d'entre elles, l'échantillonnage est dyadique en échelle, c'est-à-dire à une seule voie par octave. Des représentations à échantillonnage plus fin existent, comme celle à deux bandes par octave construite dans [Sel04]. La première section permet, à partir de représentations en ondelettes non sous-échantillonnées, de valider l'échantillonnage dyadique en échelle comme suffisant dans une perspective de description. La pyramide laplacienne, étudiée dans la deuxième section, permet d'évaluer l'impact du sous-échantillonnage spatial sur la performance de description. L'isotropie de ces représentations engendrant un sur-coût de description, les dernières sections analysent des représentations directionnelles de redondance variable.

4.1 Transformée en ondelettes non sous-échantillonnée

4.1.1 Définition

L'analyse multirésolution présentée dans la section 3.1.3 conduit à l'algorithme de Mallat permettant de construire récursivement les coefficients d'ondelettes sur une grille dyadique sous-échantillonnée. L'algorithme « à trous » [Dut89, HKMMT89] permet de calculer les coefficients d'ondelettes sur une grille dyadique non sous-échantillonnée. Pour la compression, cette représentation est inadaptée en raison de sa forte redondance, égale à $JN_b + 1$, où J est le nombre d'échelles analysées, et N_b le nombre de bandes par échelle. Elle est en revanche répandue dans de nombreux traitements visuels requérant une représentation covariante aux translations.

Pour des signaux mono-dimensionnels, le calcul des coefficients d'ondelettes sur grille non sous-échantillonnée peut s'effectuer de deux façons équivalentes. La première consiste à adapter l'algorithme de Mallat provenant de l'analyse multirésolution. À chaque itération sur le reste de basse fréquence, il suffit de calculer les convolutions avec les filtres passe-bas et passe-haut, non pas seulement aux coefficients d'indices pairs, mais également sur ceux d'indices impairs. Une façon équivalente de procéder consiste, à chaque itération, à insérer des zéros dans les filtres passe-bas h et passe-haut g utilisés dans l'algorithme de Mallat. La figure 4.2 illustre l'opération d'insertion de zéros équi-

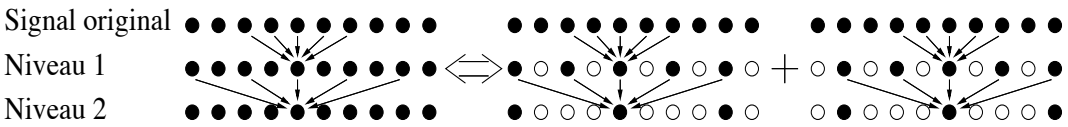

FIG. 4.2 – Équivalence entre l'algorithme à trous et l'algorithme de Mallat en considérant tous les décalages possibles avant de sous-échantillonner.

valente à l'opération de sous-échantillonnage. Les filtres h_p et g_p utilisés à la résolution p pour créer les coefficients d'approximation s_{p+1} et de détail c_{p+1} :

$$s_{p+1} = h_p \star s_p$$
$$c_{p+1} = g_p \star s_p$$

sont définis par insertion de zéros de sorte que $h_p[k] = h[k/2^k]$ si $k/2^k$ est un entier et zéro sinon. C'est pourquoi cet algorithme est connu sous le nom d'« algorithme à trous ». La transformée en ondelettes non sous-échantillonnée n'étant définie que par des convolutions, elle est bien covariante aux translations. L'algorithme de reconstruction est également similaire au cas sous-échantillonné :

$$s_p = \frac{1}{2}\left(\tilde{h}_p * s_{p+1} + \tilde{g}_p * c_{p+1}\right) \tag{4.1}$$

où les paires de filtres $(h, \tilde{h})$ et $(g, \tilde{g})$ n'ont plus à être bi-orthogonales. La condition 3.37, assurant que les recouvrements spectraux causés par le sous-échantillonnage s'annulent lors de la reconstruction, n'est plus nécessaire. La seule condition à vérifier est la condition de reconstruction parfaite 3.36. Cela conduit à une plus grande liberté dans le choix des filtres. D'autre part, la représentation n'étant pas sou-échantillonnée, l'extension aux signaux bidimensionnels peut s'effectuer avec une seule bande (isotrope) par octave, ce qui est intéressant dans une perspective de description.

L'extension aux images peut s'effectuer par produit tensoriel et conduire à trois bandes d'ondelettes par résolution. Il est également possible de se ramener à une seule bande d'ondelette par résolution en choisissant comme ondelette la fonction ψ définit par différence de fonctions d'échelle :

$$\frac{1}{4}\psi(\frac{x}{2}, \frac{y}{2}) = \phi(x, y) - \frac{1}{4}\phi(\frac{x}{2}, \frac{y}{2}) \tag{4.2}$$

Il est alors souhaitable que la fonction d'échelle ϕ et l'ondelette ψ soient isotropes. En choisissant h gaussien, et en l'approximant par un filtre binomial, les filtres d'analyse sont par exemple :

$$\begin{aligned} h_{1D} &= \frac{1}{16}[1, 4, 6, 4, 1] \\ h &= h_{1D}h_{1D} \text{ (produit tensoriel)} \qquad (4.3) \\ g &= Id - h \qquad (4.4) \end{aligned}$$

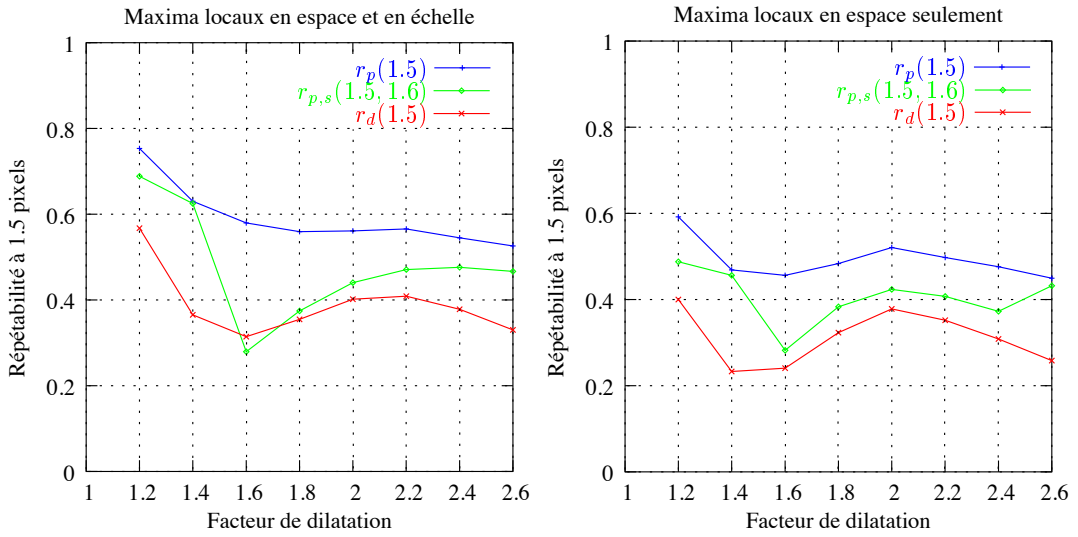

FIG. 4.3 – Représentation en ondelettes non sous-échantillonnée : répétabilités en fonction du changement d'échelle pour les maxima locaux en position et en échelle (figure de gauche), et pour les maxima locaux en position seulement (figure de droite). Les copies sont des simples dilatations.

Il existe une grande liberté dans le choix des filtres de synthèse. La reconstruction la plus évidente est donnée par $\tilde{h} = \tilde{g} = Id$. D'autres exemples de filtres de synthèse sont proposés dans [SFM].

4.1.2 Description locale

La transformée utilisée pour la description locale est isotrope. La bande d'approximation et la bande de détails sont respectivement calculées par convolution entre la bande d'approximation à la résolution précédente et les filtres g et g tels que définis par les relations 4.3 et 4.4. Les points d'intérêt sont les plus forts maxima locaux en espace et en échelle de l'énergie définie comme le carré des coefficients de la bande de détail. Leur échelle caractéristique est l'échelle à laquelle ils sont extraits. Aucune méthode d'estimation robuste d'orientations n'a été trouvée dans le domaine transformé. Une telle estimation requiert la reconstruction des bandes d'approximation. Se donnant B_p la bande de basse fréquence à la résolution p, l'orientation $\theta(\mathbf{x}, p)$ au point $\mathbf{x}$ et à la résolution p peut se définir par:

$$\theta(\mathbf{x}, p) = \text{atan2}(\frac{\partial B_p(\mathbf{x})}{\partial x}, \frac{\partial B_p(\mathbf{x})}{\partial y}) \tag{4.5}$$

où atan2 est la fonction définie par la relation 2.3 (page 58). L'énergie $E(\mathbf{x}, p)$ affectée à $\theta(\mathbf{x}, p)$ pour le calcul du descripteur SIFT est définie par :

$$E(\mathbf{x}, p) = \sqrt{\left(\frac{\partial B_p(\mathbf{x})}{\partial x}\right)^2 + \left(\frac{\partial B_p(\mathbf{x})}{\partial y}\right)^2} \tag{4.6}$$

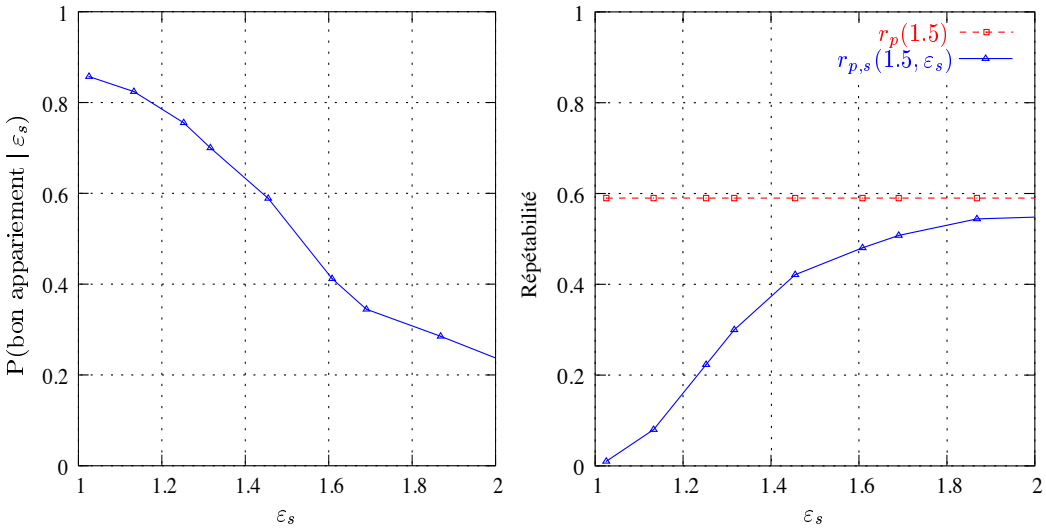

FIG. 4.4 – Représentation en ondelettes non sous-échantillonnée : sensibilité de la robustesse des caractéristiques extraites à l'erreur d'estimation en échelle. Figure de gauche : probabilité de bon appariement (au sens de la définition 11, page 46) sachant l'erreur d'estimation en échelle ε_s (telle que définie par la relation 1.25, page 44). Figure de droite : répétabilité $r_{p,s}(1.5, \varepsilon_s)$ définie par la relation 1.28.

L'orientation et l'énergie définies en tout point permettent de calculer le descripteur SIFT tel que présenté dans la section 2.4 (page 62).

La transformée en ondelettes non sous-échantillonnée ne présente pas d'intérêt pour la compression, mais permet de valider l'échantillonnage dyadique en échelle pour la détection de points et le calcul des descripteurs SIFT. Les techniques classiques utilisées en description d'images procèdent à un échantillonnage en échelle beaucoup plus fin. Cet échantillonnage grossier en échelle est commun à toutes les représentations étudiées dans cette thèse. Il est donc essentiel de valider l'échantillonnage dyadique en échelle dans une perspective de description.

Influence de l'échantillonnage dyadique en échelle sur la description. L'échantillonnage dyadique en échelle est beaucoup plus grossier que celui classiquement utilisé dans la littérature. Cela soulève deux questions.

1. Les maxima locaux en espace et en échelle sont-ils suffisamment stables pour l'extraction de points, d'échelles, et d'orientations? La figure 4.3 permet de comparer les répétabilités des maxima locaux en espace et en échelle avec celles des maxima locaux en espace seulement. Il apparaît que l'échantillonnage dyadique est encore suffisamment fin pour que les maxima locaux en échelle soient robustes. La figure montre également que les échelles et les descripteurs SIFT restent robustes dans le cas d'une discrétisation à une seule voie par octave. Pour les représentations à échantillonnage dyadique en échelle, le facteur de la « pire »dilatation est égal à

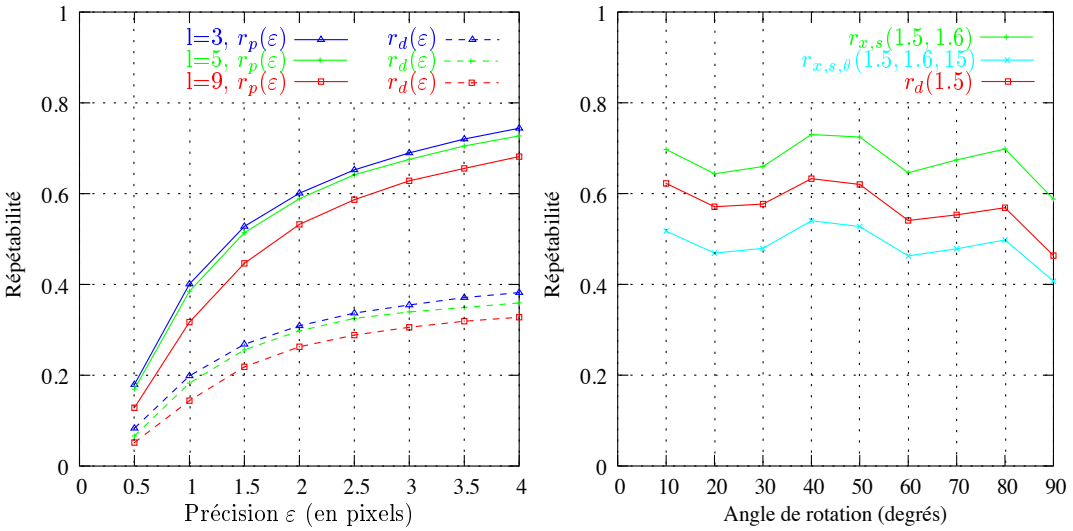

FIG. 4.5 – Représentation en ondelettes non sous-échantillonnée. Figure de gauche : répétabilités en fonction de la longueur du filtre binomial. Les copies sont créées par dilatation d'un facteur 1.6, d'une rotation de 30 degrés, d'un crop 30%, et d'une compression JPEG de facteur 10. Figure de droite : répétabilités en fonction de l'angle de rotation de l'image tournée.

1.6, et sera, dans toute la suite, utilisé pour créer les copies.

2. La sensibilité des descripteurs locaux aux erreurs d'estimation en échelle n'est-elle pas trop forte? Sur la figure 4.3, la répétabilité $r_{p,s}$ est évaluée pour une erreur de localisation de 1.5 pixels, et une erreur d'estimation en échelle $\varepsilon = 1.6$ (ε est défini par la relation 1.25, page 44). Ce choix est justifié par la figure 4.4. Sur la figure de gauche, la probabilité de bon appariement sachant $\varepsilon = 1.6$ (et sachant qu'un point distant de moins 1.5 pixels existe) est suffisante, égale à 40%. Sur la figure de droite, la probabilité qu'il existe un point extrait sur la copie à $\varepsilon \leq 1.5$ et $\varepsilon_s \leq 1.6$ est également importante, égale à 0.5.

Dans une perspective de description, un échantillonnage dyadique en échelle est donc possible.

Influence des filtres sur la répétabilité. Le choix des filtres h comme approximation de gaussienne et $g = Id - h$ permet d'approcher les coefficients d'une laplacienne à échantillonnage dyadique en échelle (à cause de l'équation de diffusion 1.12, 31). C'est pourquoi il est possible de transposer directement à ce type de représentation les techniques d'extraction de points et de description locale développées dans l'espace-échelle gaussien et présentées dans la section 2.3. En choisissant les filtres binomiaux comme approximations de la gaussienne, la figure 4.5 montre l'influence de la longueur du filtre binomial sur la robustesse des points et des descripteurs locaux. Comme dans les représentations en ondelettes à échantillonnage critique où l'ondelette de Haar conduit aux

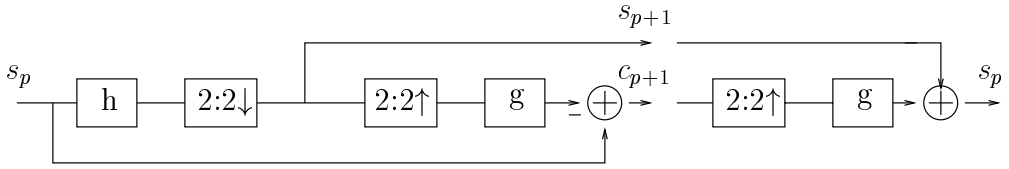

FIG. 4.6 – Un étage d'analyse (à gauche) et de synthèse (à droite) de la pyramide laplacienne.

meilleurs répétabilités, les filtres binomiaux les plus courts, donc les moins réguliers, sont mieux adaptés pour la détection de points saillants. La perte de robustesse avec la longueur du filtre est toutefois minime.

Estimation d'orientation. La méthode la plus simple d'estimation d'orientation dans le domaine transformé est de calculer l'orientation du gradient du signal transformé. Les tentatives d'estimation directement dans le domaine transformé ont échouées. Parmi les techniques d'estimation robuste d'orientation présentées dans la section 2.3, celles reposant sur le gradient peuvent facilement se transposer si l'on accepte de reconstruire les bandes d'approximation. Une telle reconstruction est peu coûteuse puisqu'elle ne requiert que N convolutions à chaque échelle, où N est le nombre de pixels. L'orientation est celle maximisant l'histogramme des orientations voisines d'après la méthode présentée dans 2.3. Les descripteurs SIFT se calculent ensuite aisément à partir de l'orientation et de l'énergie définies par les relations 4.5 et 4.6. La figure 4.5 montre la bonne robustesse des descripteurs, pour une copie créée par simple rotation (figure de droite), et pour une copie plus sévère (figure de gauche).

4.2 Pyramide laplacienne

4.2.1 Définition

Parmi les représentations multirésolution sur base fixe, seules les représentations sous-échantillonnées ont un intérêt pour la compression. La version sous-échantillonnée de la transformée en ondelettes isotrope non sous-échantillonnée est la pyramide laplacienne [BA83]. C'est une des premières et des plus simples représentations multirésolution utilisables en compression. Elle est née de la volonté d'unifier le codage prédictif et le codage basé sur les transformées décorrélantes. Elle consiste à d'abord transformer l'image en une suite d'images sous-échantillonnées et lissées par des gaussiennes. À partir d'une image grossière, il est possible de prédire l'image de résolution juste plus fine et de coder l'erreur de prédiction. La représentation est donc une pyramide d'images dont la taille est divisée par deux entre chaque résolution. Chaque image correspond à une bande de détails, sauf l'image au sommet qui est une bande d'approximation. Cette pyramide est dite laplacienne à cause de l'équation de diffusion faisant que la différence entre deux images lissées est une approximation du laplacien. La figure 4.6

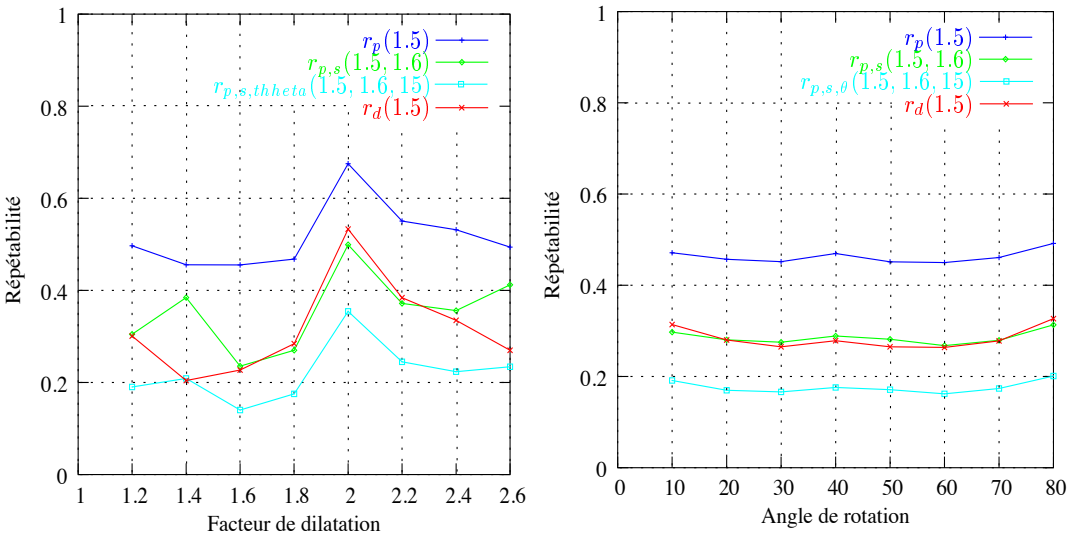

FIG. 4.7 – Pyramide laplacienne : répétabilités en fonction du changement d'échelle (figure de gauche), et en fonction de l'angle de rotation de l'image tournée (figure de droite).

montre les bancs de filtres d'analyse et de synthèse. Contrairement à la transformée discrète en ondelettes séparables, la représentation multirésolution ne conduit qu'à une seule bande (isotrope) par niveau de résolution. La bande des coefficients de détails n'est pas sous-échantillonnée, seule la bande des coefficients d'approximation l'est. La redondance de la pyramide laplacienne est suffisamment faible pour la rendre intéressante pour la compression, elle est inférieure à $\sum_{j=0}^{\infty} 4^{-j} = \frac{4}{3}$. Comme dans l'algorithme de Mallat, les coefficients d'approximation $s_p[n_x, n_y]$ à l'échelle p se calculent à partir de ceux à l'échelle précédente par convolution et décimation :

$$s_p[n_x, n_y] = \sum_{k_x, k_y \in \mathbb{Z}} h[2n_x - k_x, 2n_y - k_y] s_{p-1}[k_x, k_y] \tag{4.7}$$

Les coefficients de détails $c_p[n_x, n_y]$ constituent une erreur de prédiction :

$$c_p[n_x, n_y] = s_p[n_x, n_y] - \sum_{k_x, k_y \in \mathbb{Z}} g[2n_x - k_x, 2n_y - k_y] s_{p-1}[k_x, k_y] \tag{4.8}$$

La reconstruction la plus simple est donnée par :

$$s_p[n_x, n_y] = c_p[n_x, n_y] + \sum_{k_x, k_y \in \mathbb{Z}} g[2n_x - k_x, 2n_y - k_y] s_{p-1}[k_x, k_y] \tag{4.9}$$

La reconstruction optimale, au sens de la minimisation de l'erreur quadratique en présence de bruit blanc uniformément réparti dans chacune des sous-bandes, est de complexité légèrement supérieure, et est présentée dans [DV03].

4.2.2 Description locale

Le sous-échantillonnage rend la pyramide laplacienne variante aux translations et aux rotations. Néanmoins, le sous-échantillonnage n'a lieu que dans la bande d'approximation, ce qui limite considérablement cette variance. La transposition de la description SIFT dans les pyramides laplaciennes est présentée dans cette section.

Détection de points d'intérêt, et d'échelles caractéristiques. L'extraction de maxima locaux en espace et en échelle devient délicate à cause des grilles de résolution variable en fonction de l'échelle. Les meilleurs résultats ont été obtenus par extraction des maxima locaux en espace seulement, et en les localisant sur la grille de pleine résolution par la technique de Loupias [LS99] présentée dans la section 3.3.2. L'échelle caractéristique d'un point d'intérêt est celle où l'énergie normalisée en échelle est maximale à travers les échelles. En comparant les figures 4.3 et 4.7 donnant les répétabilités des caractéristiques extraites à partir des représentations en ondelettes non sous-échantillonnées et à partir des pyramides laplaciennes, il apparaît que la dégradation causée par le sous-échantillonnage est au pire de 50% (pour les rotations et les changements d'échelle de petit facteur). Le taux d'appariements corrects à 1.5 pixels défini par la relation 1.32 (page 46) est, pour le type de copies utilisé par la figures 3.17 (page 97) de $0.2/0.55 \simeq 0.36$ pour la transformée en ondelettes à échantillonnage critique. La figure 4.8 montre que ce taux devient $0.22/0.45 \simeq 0.5$ pour la pyramide laplacienne. Une redondance faible permet donc de sensiblement améliorer la robustesse des caractéristiques extraites. Enfin, il est important de souligner que lorsque le facteur de dilatation correspond au facteur de sous-échantillonnage (égal à deux dans toutes les représentations étudiées par la suite), les représentations de l'image originale et de la copie se correspondent. Ceci explique le rebond important que marquent les répétabilités au facteur de dilatation égal à deux.

Estimation d'orientation, et description des points d'intérêt. L'isotropie des représentations par pyramide laplacienne implique un coût de description. Comme pour les représentations en ondelettes isotropes non sous-échantillonnées, les tentatives d'extraction robuste d'orientation directement dans le domaine transformé ont échoué. Une telle extraction est possible au prix de la reconstruction des bandes d'approximation à chaque échelle. Le coût est faible, il est de N convolutions et de N additions par échelle, où N est le nombre de coefficients à une échelle donnée. L'orientation est alors estimée à partir de l'orientation du gradient des bandes reconstruites. L'énergie permettant de pondérer le poids des orientations est la norme du gradient. Les figures 4.7 et 4.8 montrent que les orientations ainsi extraites, et les descripteurs SIFT calculés à partir de ces orientations, sont très robustes. La pyramide laplacienne sera donc une des représentations choisies dans dans les schémas de compression présentés dans le dernier chapitre. Le reste du chapitre analyse des représentations directionnelles, de sorte à disposer de l'information d'orientation sans avoir à reconstruire l'image.

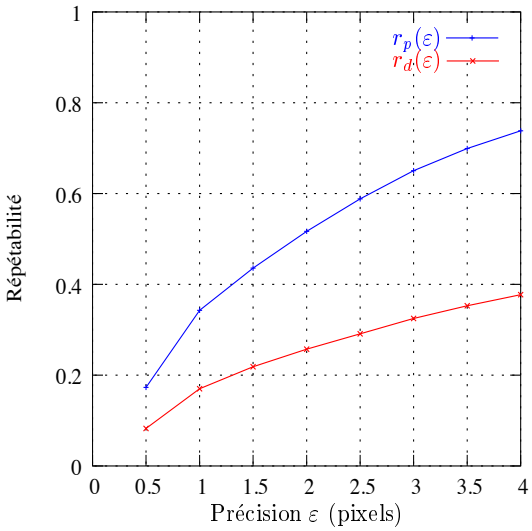

FIG. 4.8 – Pyramide laplacienne : répétabilités en fonction de la précision de localisation des points extraits, pour une copie synthétisée par dilatation de facteur 1.6, d'une rotation de 30 degrés, d'une compression JPEG de facteur 10, et d'un crop de facteur 30.

4.3 Transformée en contourlets

4.3.1 Définition

La représentation en contourlets se situe dans le prolongement des travaux de Candès et Donoho [CD99] visant à adapter les représentations en ondelettes aux spécificités des images naturelles.

Sous-optimalité des représentations en ondelettes séparables. Les images naturelles peuvent être considérées comme appartenant à l'espace de fonctions de $L^2(\mathbb{R}^2)$ régulières partout sauf sur un ensemble de courbes elles aussi régulières (les contours). Les coefficients d'ondelettes séparables ne peuvent pas prendre en compte la régularité des contours et sont fortement corrélés en valeur absolue le long des contours. Une façon de mesurer l'adaptation d'une représentation aux images naturelles est de considérer la décroissance de l'erreur entre l'image originale I et l'image I_m reconstruite à partir des m plus grands coefficients (aussi appelée approximation non linéaire à m termes). Pour les images régulières de classe C^p, cette décroissance vaut dans le cas des ondelettes séparables :

$$||I - I_m||^2 \leq Cm^{-p} \text{ où } C \text{ est une constante} \tag{4.10}$$

Pour les images de classe C^p partout sauf sur un ensemble de courbes régulières, la décroissance devient :

$$||I - I_m||^2 \leq Cm^{-1} \tag{4.11}$$

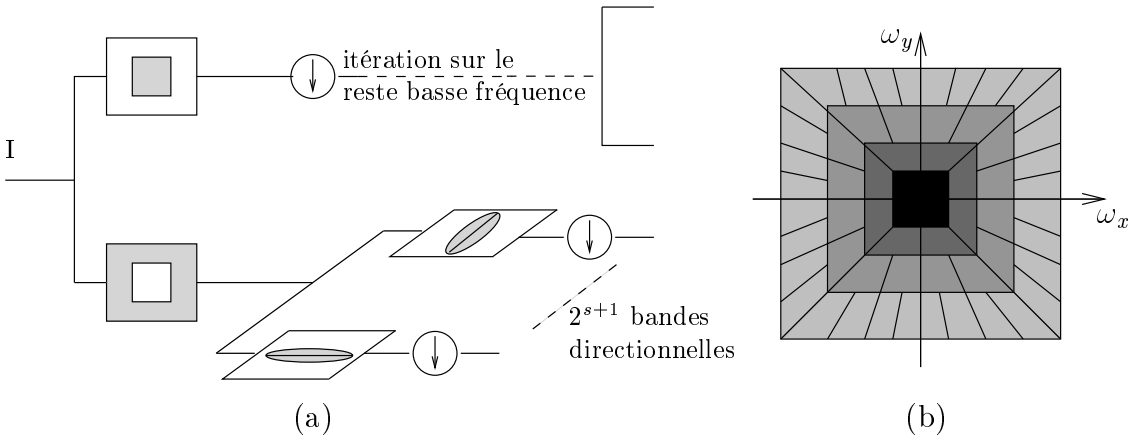

FIG. 4.9 – À gauche : bancs de filtres utilisés dans une transformée en contourlets. Un banc de filtres directionnels sur la bande haute d'une pyramide laplacienne, et le schéma est itéré sur la bande basse. À droite : partition du plan fréquence associée au schéma de gauche.

montrant ainsi que les ondelettes séparables ne sont pas optimales pour la représentation des images naturelles.

Transformée en curvelets. Cette transformée permet d'obtenir une meilleure approximation non linéaire que celle des ondelettes séparables. Elle consiste à décomposer l'image par une pyramide laplacienne, à partitionner chacune des sous-bandes en blocs de taille liée à la fréquence centrale de la sous-bande, puis à appliquer une transformée en ridgelets sur chacun des blocs. Une ridgelet $\psi_{a,s,\theta}$ est générée à partir d'une ondelette mono-dimensionnelle ψ par :

$$\psi_{a,s,\theta}(x,y) = s^{-1/2}\psi(\frac{x\cos(\theta)+y\sin(\theta)}{s}) \tag{4.12}$$

La décomposition par la pyramide laplacienne permet de partitionner l'image lissée en blocs de tailles croissantes, et donc d'obtenir par la transformée en ridgelets la taille maximale pour laquelle le contour est approximativement une droite. Il est montré dans [CD99] que la décroissance de l'erreur pour les images de classe C^2 partout sauf sur un ensemble de courbes régulières est :

$$||I - I_m||^2 \leq Cm^{-2}\log_2(m^3) \tag{4.13}$$

mettant en évidence l'avantage d'une analyse directionnelle sur l'analyse par filtres séparables. Il existe néanmoins deux limitations importantes à la transformée en curvelets.

1. Des artefacts apparaissent dans l'image reconstruite à cause de la partition des sous-bandes en blocs.

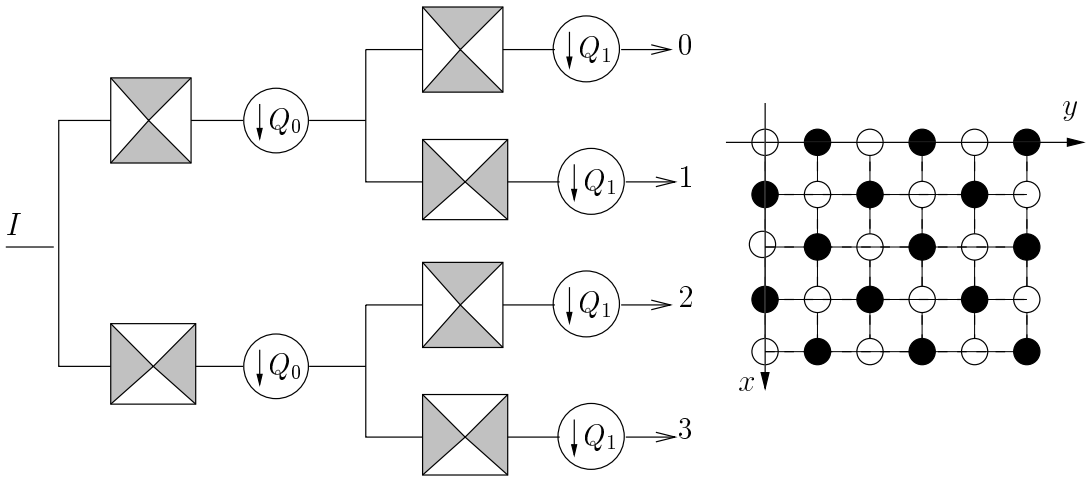

FIG. 4.10 – À gauche : banc de filtres directionnels dans une transformée en contourlets. L'analyse directionnelle est effectuée par convolution et décimation sur grille quinconce. À droite : les deux grilles associées à un sous-échantillonnage quinconce.

2. La transformée est définie en coordonnées polaires, ce qui créé une difficulté d'implémentation pour des images discrètes à support rectangulaire. La première implémentation [SCD02], destinée au débruitage, conduit à une représentation en curvelets de redondance $16J + 1$ où J est le nombre de niveaux de résolution, ce qui est prohibitif pour une application en compression.

Transformée en contourlets. La transformée en contourlets proposée par Do dans [Do01] conserve la pyramide laplacienne comme premier banc de filtres multirésolution, mais utilise un deuxième banc de filtres directionnels pour grouper les coefficients situés sur des contours. La figure 4.9 représente ces deux bancs de filtres mis en cascade. Le nombre de bandes directionnelles est une puissance de deux ; il est fixé sur la figure à 2^{s+1}, où s est l'échelle analysée, conduisant à la partition du plan fréquence illustrée sur la figure de droite.
L'analyse directionnelle s'obtient par la mise en cascade d'un banc de filtres à deux canaux tel qu'illustré par la figure 4.10 de gauche. Chaque étage de la décomposition est constitué d'un filtrage horizontal ou vertical suivi d'un sous-échantillonnage sur une grille quinconce. Les deux grilles associées aux sous-échantillonnages quinconces Q_1 et Q_2 sont respectivement constituées de l'ensemble des points noirs et blancs de la figure 4.10 de droite. La grille de $\mathbb{Z}^2$ associée à une matrice $M \in \mathcal{M}_2(\mathbb{Z}^2)$ étant définie par :

$$Grille(M) = \{\mathbf{x} : \mathbf{x} = M\mathbf{n},\ \mathbf{n} \in \mathbb{Z}^2\} \tag{4.14}$$

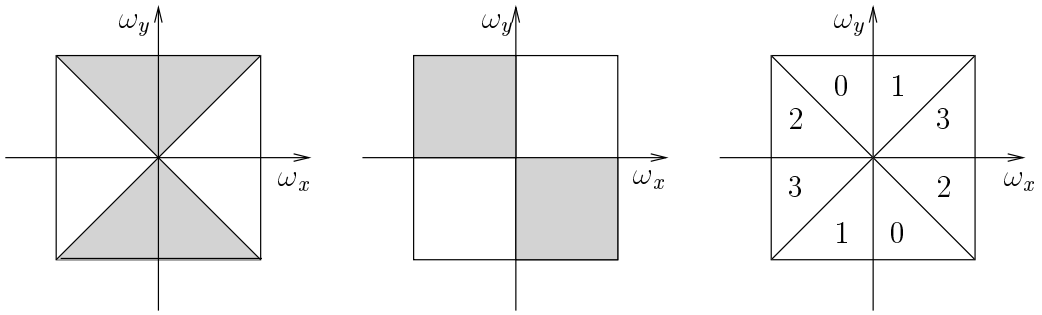

FIG. 4.11 – Filtre horizontal (à gauche), filtre quadrant obtenu par composition d'un filtre horizontal et d'un sous-échantillonnage quinconce (au milieu), et partition du plan fréquence obtenue par composition de filtres horizontaux, verticaux, et quadrants.

les matrices Q_1 et Q_2 associées aux sous-échantillonnages quinconces sont égales à :

$$Q_1 = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \quad Q_2 = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \tag{4.15}$$

La combinaison astucieuse de filtres horizontaux et verticaux avec des sous-échantillonnages quinconces permet d'obtenir le filtre quadrant représenté au milieu de la figure 4.11. Il s'agit du filtre équivalent au niveau du deuxième étage du banc de filtres directionnels de la figure 4.10. En combinant les filtres horizontaux et verticaux avec ces filtres quadrants, il est possible d'obtenir la partition du plan fréquence illustré à droite de la figure 4.11. Le problème du filtrage directionnel à échantillonnage critique est l'instabilité des orientations détectées.

4.3.2 Description locale

Le banc de filtres directionnels à échantillonnage critique permet de créer une représentation creuse et de faible redondance. Les représentations en contourlets sont intéressantes en compression, mais pas en description pour deux raisons :

- le sous-échantillonnage quinconce permettant l'analyse directionnelle conduit à des bandes de taille variable, rendant ainsi difficile la localisation spatiale relative entre deux bandes directionnelles ;
- l'échantillonnage critique du banc de filtres directionnels ne permet pas une estimation robuste d'orientations.

Le premier problème conduit soit à extraire les maxima locaux indépendamment dans chaque bande, soit à définir le voisinage d'un coefficient dans les autres bandes directionnelles comme étant de taille correspondant à l'incertitude de localisation, c'est-à-dire de taille 2×2 coefficients, soit à interpoler chacune des bandes sur une grille de taille double. Les deux dernières solutions conduisent à des répétabilités similaires en terme de points, d'échelles, d'orientations et de descripteurs. Notant $B_{p,k}$ la $k^{\text{è}}$ bande directionnelle ($1 \leq k \leq N$) à la résolution p, l'extraction la moins complexe conduit aux

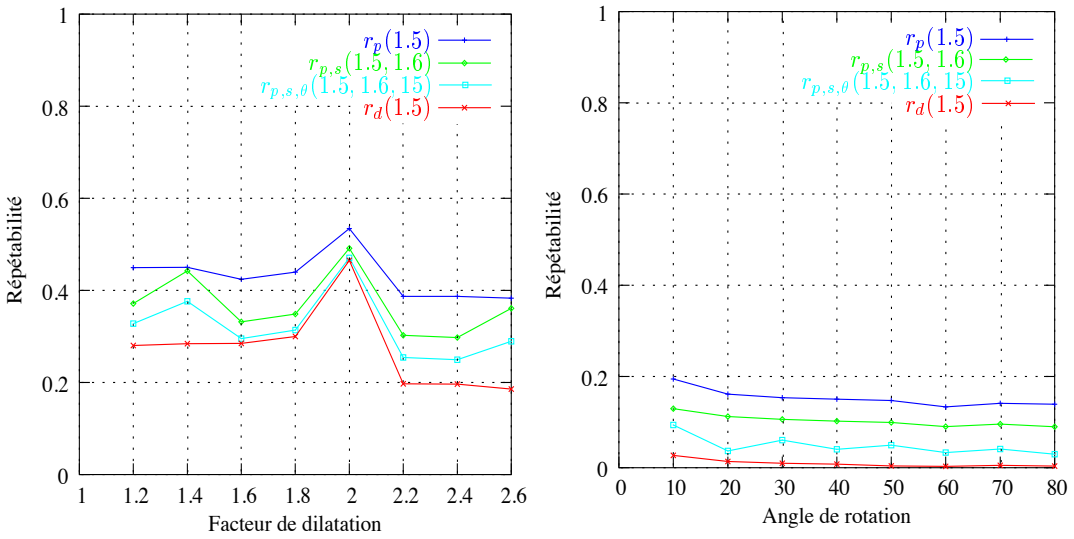

FIG. 4.12 – Représentation en contourlets : robustesse des caractéristiques face aux dilatations (figure de gauche) et aux rotations (figure de droite).

points d'intérêt $\mathbf{x}$ de la forme:

$$\mathbf{x} = \{\mathbf{x} \in B_{p,k} \,|\, \forall 1 \leq j \leq N, \forall \mathbf{y} \in V_{kj}(\mathbf{x}), |B_{p,k}(\mathbf{x})| > |B_{p,j}(\mathbf{y})|\} \tag{4.16}$$

où, k étant fixé, $V_{kk}(\mathbf{x})$ est le voisinage contenant les 8 plus proches voisins de $\mathbf{x}$ dans la bande $B_{p,k}$, et les $\{V_{k,j}\}_{j \neq i}$ sont les points des autres bandes directionnelles couvrant toute l'incertitude de localisation de $\mathbf{x}$ dans la bande j. Considérant un point $\mathbf{x}$ extrait dans la bande $B_{p,k}$, son échelle s et son orientation θ caractéristiques sont simplement données par:

$$\begin{aligned} s &= 2^p \\ \theta &= (-1)^{\frac{B_{p,k}(\mathbf{x})}{|B_{p,k}(\mathbf{x})|}} \frac{(k-1)\pi}{N} \end{aligned}$$

L'énergie associée à l'orientation θ pour le calcul du descripteur SIFT est $B_{p,k}(\mathbf{x})^2$. La figure 4.3.2 donne les répétabilités des caractéristiques ainsi extraites face aux dilatations et aux rotations. La comparaison des résultats avec ceux obtenus avec une pyramide laplacienne permet de conclure sur la performance d'une analyse directionnelle à échantillonnage critique dans une perspective de description.

- Face aux dilatations, ou à toute transformation ne mettant pas en jeu de rotations, la robustesse des points et des échelles n'est pas modifiée par l'analyse directionnelle. Les orientations extraites sont en revanche plus robustes, conduisant à une meilleure répétabilité par descripteurs r_d.
- Face aux rotations, les caractéristiques extraites sont très instables. Cette variance de la représentation aux rotations est observée quelque soit le nombre de bandes directionnelles. Comme la variance aux translations des représentations en ondelettes du précédent chapitre, elle provient de l'échantillonnage critique du banc

de filtre directionnel. Un fort recouvrement spectral a lieu dans les bandes directionnelles, si bien qu'elles portent de l'énergie correspondant à des orientations d'autres bandes.

La conclusion de l'analyse des représentations par contourlets est qu'une analyse directionnelle par un banc de filtres à échantillonnage critique ne permet pas d'extraire des orientations robustes. Le sous-échantillonnage introduit un recouvrement spectral faisant que chacune des bandes orientées contient de l'énergie correspondant aux directions des autres bandes. Le reste du chapitre analyse donc des bancs de filtres directionnels redondants.

4.4 Transformée en ondelettes complexes

4.4.1 Définition

Une représentation multirésolution covariante aux translations, de redondance égale à quatre, et à six bandes directionnelles par octave, a été proposée par Kingsburry à partir d'ondelettes complexes [Kin98]. Une transformée est covariante aux translations si le signal reconstruit à partir de chacune des sous-bandes l'est. Pour les transformées multirésolution échantillonnées, l'expression de ce signal reconstruit à partir des matrices polyphases permet de mettre en évidence les réplications spectrales causées par le sous-échantillonnage comme la source de la variance aux translations. La condition 3.41 (page 87) est une caractérisation spectrale des transformées multirésolution échantillonnées covariantes aux translations. Il a été montré dans la section 3.2.1 que la principale violation de cette condition par les ondelettes réelles est due à la composante spectrale non nulle dans les fréquences négatives, qui après réplication recouvre fortement la composante positive. Pour y remédier, Kingsburry propose des ondelettes complexes à spectre analytique, c'est-à-dire dont la transformée de Fourier est nulle sur un demi-plan [Kin98]. Cette transformée présente de nombreux avantages pour le problème conjoint de compression et de description.

- L'implémentation proposée dans [Kin98] est très rapide, et s'effectue au moyen de deux bancs de filtres en parallèle.
- La redondance de la transformée est égale à quatre, ce qui est acceptable pour des applications en compression. Dans [KR03], des projections sur ensembles convexes permettent d'obtenir des taux de compression voisins de JPEG2000.
- La représentation est presque covariante aux translations.
- L'extension aux signaux bidimensionnels s'effectue par produit tensoriel. Considérant trois ondelettes mono-dimensionnelles à spectre analytique, leurs produits tensoriels partitionnent le premier quadrant du plan espace-fréquence en trois. Trois autres ondelettes analytiques sont nécessaires pour partitionner le second quadrant, et ainsi disposer de toute l'information pour reconstruire des images réelles. Les ondelettes complexes ainsi créées par produit tensoriel sont orientées et permettent l'analyse directionnelle utile pour la description (contrairement à

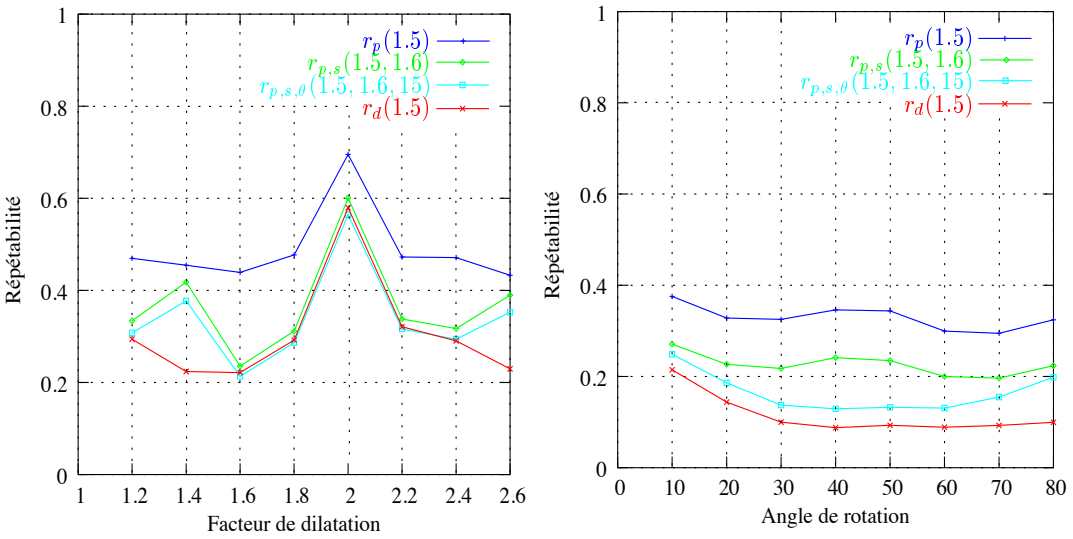

FIG. 4.13 – Représentation en ondelettes complexes : robustesse des caractéristiques face aux dilatations (figure de gauche) et aux rotations (figure de droite).

la bande diagonale apparaissant pour des ondelettes réelles).

4.4.2 Description locale

La transformée en ondelettes complexes s'effectue par l'analyse récursive des bandes de basse fréquence par trois filtres passe-haut complexes en quadrature, et par un filtre passe-bas complexe. La représentation est donc constituée de six bandes (réelles) directionnelles par résolution et d'un reste (complexe) de basse fréquence. Les caractéristiques les plus robustes ont été extraites de la manière suivante:

- les points d'intérêt sont les maxima locaux (en position seulement) de l'énergie définie comme le plus fort carré des 6 coefficients réels correspondant aux bandes orientées ;
- les points d'intérêt sont recalés sur la grille la plus fine (de résolution quatre fois plus petite que l'image) par la méthode de Loupias, ce qui permet d'extraire l'échelle caractéristique comme l'échelle de plus forte énergie à travers les échelles ;
- l'orientation estimée à partir de la relation 2.6 (page 61). Ceci est possible car les filtres sont complexes en quadrature. L'orientation est donc égale à la moitié de la phase du complexe $\mathbf{z}$ défini par :

$$\mathbf{z} = \sum_{k=1}^{3} c_k e^{i\frac{2(k-1)\pi}{3}} \tag{4.17}$$

où c_k est le coefficient complexe dans la k-ème bande orientée selon $\frac{(k-1)\pi}{3}$.

La robustesse des caractéristiques ainsi extraites est bonne pour les dilatations, mais décevantes pour les rotations. Le problème apparaît principalement dans la robustesse des points et des échelles. Les orientations sont en effet d'une bonne robustesse. Le

problème peut provenir du fait que, malgré la redondance de la représentation égale à 4, la grille la plus fine est 4 fois plus petite que l'image de pleine résolution, contrairement à la pyramide laplacienne où la bande de détails n'est pas sous-échantillonnée. La prochaine section analyse les représentations orientables, qui sont constituées d'une pyramide laplacienne, suivie d'une analyse directionnelles non sous-échantillonnée.

4.5 Transformées orientables

4.5.1 Définition

Interpolation de fonctions échantillonnées. Dans la transformée en ondelettes à échantillonnage critique, les filtres passe-bas et passe-haut n'étant pas idéaux, la fréquence d'échantillonnage dans chaque sous-bande est inférieure à la fréquence de Nyquist. En utilisant des bandes orthogonales, il est néanmoins possible de reconstruire parfaitement le signal. Il n'est, en revanche, pas possible d'interpoler les coefficients transformés aux positions échantillonnées. Dans le cas de signaux continus, un problème classique est de savoir dans quelles conditions la convolution d'un signal $f(x)$ par un noyau $h(x)$ peut être échantillonnée sans perte d'information. La fréquence minimale d'un tel échantillonnage est la fréquence de Nyquist égale à la bande-passante du noyau h. Dans le cas où le signal $f(x)$ admet une décomposition en série de Fourier (en particulier les signaux périodiques ou à support compact), une formulation analytique des fonctions d'interpolation est donnée dans [FA91]. Considérant sans restriction que le signal est périodique de période 2π, les N coefficients $y[n]$ uniformément échantillonnés s'écrivent :

$$y[n] = \int_0^{2\pi} f(x)h(n\Delta_x - x)dx \tag{4.18}$$

où Δ_x est le pas d'échantillonnage. La transformée est dite « translatable »(plus connue sous le nom anglais « shiftable ») s'il existe des fonctions d'interpolation $b_n(x_0)$ permettant de reconstruire le signal en tout point x_0 :

$$\forall x_0,\ f(x_0) = \int_0^{2\pi} f(x)h(x_0 - x)dx = \sum_{n=0}^{N-1} b_n(x_0)y[n] \tag{4.19}$$

Pour que cette relation soit vraie pour tout signal $f(x)$, le noyau h doit vérifier :

$$\forall x_0,\ h(x_0 - x) = \sum_{n=0}^{N-1} b_n(x_0)h(n\Delta_x - x) \tag{4.20}$$

Le signal f étant périodique, le noyau h peut également être considéré périodique. Notant $h(x) = \sum_{k=0}^{N-1} H(k)e^{2j\pi\frac{k}{N}}$ la série de Fourier de h, la relation précédente s'écrit :

$$\forall x_0, k,\ H(k)e^{jkx_0} = H(k)\sum_{n=0}^{N-1} b_n(x_0)e^{jkn\Delta_x} \tag{4.21}$$

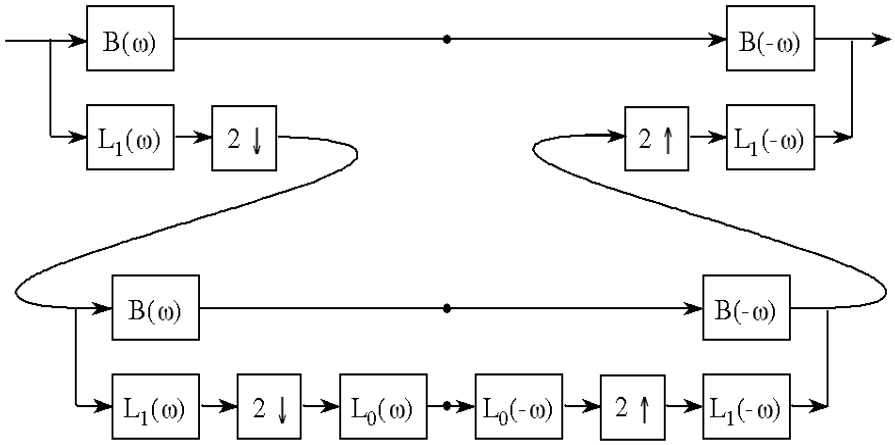

FIG. 4.14 – Premier banc de filtres utilisé dans une représentation orientable. Un deuxième banc de filtres directionnels est appliqué sur chacune des bandes hautes.

Deux noyaux h partageant le même ensemble $\{k_0, k_1, \ldots, k_{M-1}\}$ de fréquences pour lesquelles H est non nulle, ont donc le même ensemble de fonctions d'interpolation.

Conséquences pour la description. Deux résultats sont très importants pour le problème de description, et sont respectivement démontrés dans [SAH92] et [FA91].

- Le premier résultat est une caractérisation des transformées « translatables », permettant d'obtenir des représentations strictement covariantes aux translations. Une transformée multirésolution est « translatable »si l'énergie globale $\sum_{n=0}^{N-1} |y[n]|^2$ est invariante aux translations (ce qui est équivalent à un recouvrement spectral inter-bandes nul).
- Le second résultat est la possibilité de construire des filtres orientables à partir de la relation 4.21. Un filtre est orientable si la composante polaire de sa transformée de Fourier est « translatable ». Un exemple simple est donné par le filtre orienté selon θ_0 et défini par $h_{\theta_0}(\theta, r) = cos(\theta - \theta_0)c(r)$, où $c(r)$ est un filtre symétrique quelconque. Des relations trigonométriques standards montrent en effet que la réponse impulsionnelle du filtre h_{θ_0} peut être interpolée à partir des réponses des filtres h_0 et $h_{\frac{\pi}{2}}$ par

$$h_{\theta_0} = \cos(\theta_0)h_0 + \sin(\theta_0)h_{\frac{\pi}{2}}$$

  La conception de filtres orientables a été introduite dans [FA91].

Implémentation. Dans [SAH92] est proposée une implémentation par banc de filtres d'une transformée multirésolution sous-échantillonnée « translatable »et orientable. La transformée est séparable dans le domaine de Fourier et définie par $T(\omega, \theta) = U(\omega)H(\theta)$. Un premier banc de filtres décompose le signal en une bande haute sur laquelle opère

un second banc de filtres orientables, et une bande basse sur laquelle est itérée la décomposition. Pour que la transformée soit « translatable », il ne doit pas y avoir de recouvrements spectraux inter-bandes. Cela peut s'approcher en utilisant comme premier banc de filtres la pyramide laplacienne illustrée par la figure 4.14, où la bande haute est filtrée par un passe-bande, et la bande basse par un passe-bas puis sous-échantillonnée. La réponse du système s'écrit :

$$S(\omega) = |B(\omega)|^2 + |L_1(\omega)|^2 |L_0(2\omega)|^2 \tag{4.22}$$

Comme $B(\omega)$ est passe-bande et les $L_i(\omega)$ passe-bas, $S(\omega)$ est passe-bas. La reconstruction du signal requiert donc le calcul d'un reste passe-haut

$$R(\omega) = 1 - S(\omega) \tag{4.23}$$

Pour mettre en cascade le système selon le schéma de la figure 4.14, la réponse du système $S(\omega)$ doit être égale à $|L_0(\omega)|^2$, conduisant à :

$$|L_0(\omega)|^2 = |B(\omega)|^2 + |L_1(\omega)|^2 |L_0(2\omega)|^2 \tag{4.24}$$

Le second banc de filtres décompose $B(\omega)$ en N bandes directionnelles orientables. En choisissant comme réponse directionnelle

$$H(\theta) = \cos^3(\theta) = \frac{1}{4}\cos(3\theta) + \frac{3}{4}\cos(\theta) \tag{4.25}$$

les fonctions d'interpolations, données par la relation 4.21, sont :

$$b_n(\theta) = \frac{1}{2}\left(\cos(\theta - n\Delta_\theta) + \cos(3(\theta - n\Delta_\theta))\right),\ 0 \leq n \leq 3 \tag{4.26}$$

où $\Delta_\theta = \frac{\pi}{4}$. La redondance d'une représentation orientable sur J octaves à n bandes directionnelles par octave est élevée et vaut $1 + \sum_{j=0}^{J-1} \left(\frac{1}{4}\right)^j \simeq 1 + \frac{4}{3}n$.

4.5.2 Description locale

Le premier banc de filtres décrit par la figure 4.14 est fixé. La réponse impulsionnelle $L_1(\omega)$ est celle d'un filtre binomial de longueur 7, et la réponse $L_0(\omega)$ est construite pour être unitaire entre 0 et $\frac{\pi}{2}$ et nulle en π par un filtre de longueur 13. Le deuxième banc de filtres décompose la bande de haute fréquence du premier banc de filtres à partir de filtres dont la composante polaire de la transformée de Fourier est de la forme $H_k(\theta) = \cos^k(\theta)$.

Notant $B_{p,n}$ la $n^{\text{è}}$ bande directionnelle ($1 \leq n \leq N$), et b_n la $n^{\text{è}}$ fonction d'interpolation, l'énergie au point $\mathbf{x}$ à l'échelle p est définie par :

$$E_p(\mathbf{x}) = \max_{0 \leq \phi < 2\pi} \sum_{1 \leq n \leq N} b_n(\alpha) B_{p,n}(\mathbf{x}) \tag{4.27}$$

et l'orientation par :

$$\theta_p(\mathbf{x}) = \arg \max_{0 \leq \phi < 2\pi} \sum_{1 \leq n \leq N} b_n(\alpha) \tag{4.28}$$

Les meilleurs résultats ont été obtenus en extrayant les plus forts maxima locaux de l'énergie E_p ainsi définie, et en recalant les points sur la grille de pleine résolution par la méthode de Loupias. L'échelle d'un point est l'échelle de la plus énergie à travers les échelles. L'orientation et l'énergie permettent de transposer directement la description SIFT présentée dans la section 2.4.

Détermination du nombre de bandes directionnelles. Il est montré dans [FA91] que le nombre minimal de bandes directionnelles pour construire la réponse $H_k(\theta)$ du deuxième banc de filtres orientables est k. La sélectivité angulaire croît avec le nombre de bandes. La figure 4.15 montre que la robustesse des orientations extraites au niveau des points d'intérêt n'est pas améliorée avec le nombre de bandes directionnelles. Le nombre de bandes choisi pour le problème conjoint de compression et de description est donc égal à 2, de sorte à minimiser la redondance de la représentation.

Évaluation d'une représentation orientable à deux bandes. La figure 4.16 montre les bonnes répétabilités pour les dilatations, et pour une transformation sévère composée d'une dilatation, d'une rotation, d'un crop, et d'une compression JPEG. Il est intéressant de comparer cette figure avec la figure 4.8 donnant les répétabilités pour le même type de copie, mais pour des caractéristiues (points et descripteurs SIFT) extraites à partir de pyramide laplacienne. Le premier banc de filtres utilisé dans les représentations orientables est également une pyramide laplacienne. Le deuxième banc de filtres, non sous-échantillonné, ne dégradent que faiblement la robustesse des points extraits : la répétabilité à 1.5 pixels est de 0.45 pour les pyramides laplaciennes, et de 0.4 pour les pyramides orientables. L'estimation d'orientation est, en revanche, plus coûteuse en temps de calcul (elle nécessite le calcul des bandes d'approximation) et moins fiable dans les pyramides laplacienne que dans les pyramides orientables. À une précision de 1.5 pixels, la répétabilité par descripteur le plus proche passe de 0.22 dans les pyramides laplaciennes à 0.28 dans les pyramides orientables.

4.6 Conclusion

Des transformées multirésolution redondantes, isotropes ou directionnelles, ont été analysées dans un but de description locale. Les bonnes performances de description dans les représentations en ondelettes non sous-échantillonnées permettent de valider l'échantillonnage dyadique en échelle. Les techniques classiques de description locale utilisent un échantillonnage beaucoup plus fin en échelle pour un gain pouvant être faible. La pyramide laplacienne est une représentation isotrope et requiert la reconstruction de l'image pour l'estimation de l'orientation. Elle est néanmoins intéressante pour le

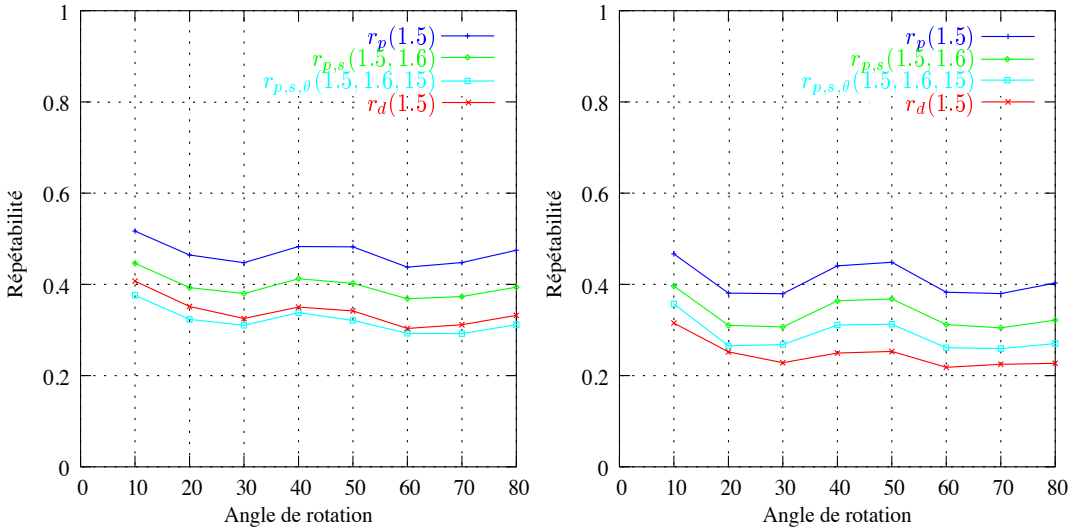

FIG. 4.15 – Représentation orientables : robustesse des caractéristiques face aux rotations, pour deux bandes directionnelles (figure de gauche), et quatre bandes directionnelles (figure de droite).

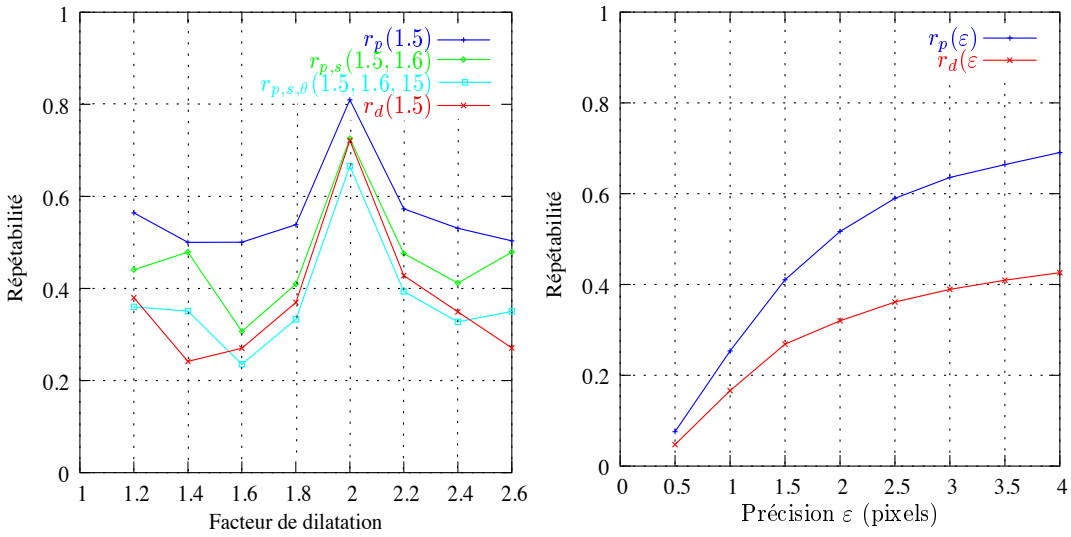

FIG. 4.16 – Représentations orientables : répétabilités face aux dilatations (figure de gauche), et face à une transformation plus sévère composée d'une dilatation de facteur 1.6, d'une rotation d'angle 30 degrés, d'un crop de facteur 30%, et d'une compression JPEG de facteur 10.

problème conjoint de compression et de description, car elle est de faible redondance et conduit à de bonnes répétabilités des caractéristiques extraites. La pyramide laplacienne valide ainsi la possibilité de sous-échantillonner dans un but de description locale. L'investigation de représentations directionnelles conduit à des résultats assez décevants. De telles représentations sont potentiellement très intéressantes car offrent la possibilité de décrire pour un coût très faible : les informations d'orientation et d'énergie sont directement accessibles. Les bancs de filtres directionnels à échantillonnage critique, comme les contourlets, conduisent à des représentations trop variantes aux rotations pour être utilisables en description. Les ondelettes complexes donnent de bons résultats en estimation robuste d'orientation, mais pas en détection robuste de points et d'échelle, à cause de la grille la plus fine qui n'est que de taille le quart de celle de l'image originale. Les représentations orientables conduisent à de bons résultats pour tout type de transformation admissible, mais sont de redondance élevée. Il existe néanmoins des techniques de codage pour ce type de représentations. Le chapitre suivant propose des schémas de compression adaptés à la description locale.

Chapitre 5

Effets de la compression sur la description locale

Les chapitres précédents ont porté sur la première étape du schéma de compression, à savoir la transformée d'images adaptée au problème conjoint de compression et de description. Ces transformées sont inversibles, et la perte d'information n'apparaît que dans les étapes de quantification et éventuellement de codage. Le but de ce chapitre est de mesurer la dégradation sur la description locale causée par cette perte d'information. Deux schémas de compression sont proposés dans un but de description locale. Le premier repose sur une représentation de faible redondance, la pyramide laplacienne ; le second sur une représentation de forte redondance, la représentation orientable. La dernière étape du schéma de compression, à savoir le codage des coefficients quantifiés, n'est pas étudié. Le problème de concevoir un codage adapté à la description est en effet très ardu et n'a jamais été abordé. La solution proposée est de coder les coefficients quantifiés par un codeur entropique de complexité linéaire. Les performances de compression sont donc données en terme de PSNR et d'entropie marginale, et les performances de description en terme de répétabilités des caractéristiques extraites.

5.1 Codage des pyramides laplaciennes

5.1.1 Impact du bruit de quantification sur la description

La distribution des coefficients de la pyramide laplacienne d'une image naturelle suit approximativement la loi gaussienne généralisée définie par la relation 3.32 (page 82). Une quantification uniforme de chacune des bandes est alors proche de la quantification scalaire optimale au sens de la minimisation de l'erreur quadratique. Les quantifications vectorielles permettent d'obtenir de meilleures performance, mais la création du dictionnaire des vecteurs de quantification est un problème ardu si l'on doit prendre en compte les contraintes liées à la description. La figure 5.1 donne la répétabilité et le PSNR en fonction de l'entropie des pyramides laplaciennes uniformément quantifiées. L'inversion

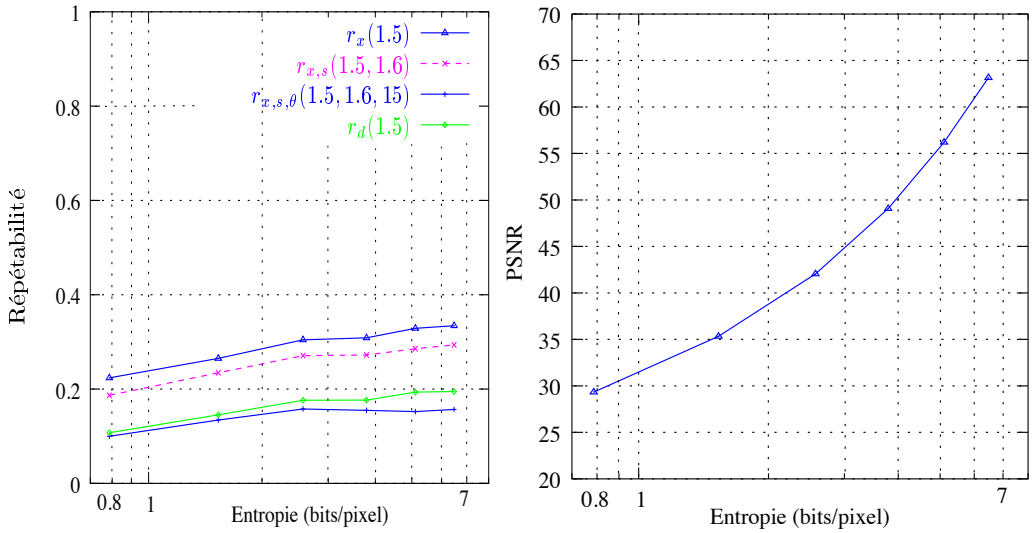

FIG. 5.1 – Répétabilités et PSNR en fonction de l'entropie. Les coefficients de la basse fréquence sont quantifiés sur 8 bits, et ceux des bandes hautes entre 3 et 7 bits.

de la pyramide, nécessaire pour le calcul du PSNR, est décrite par la figure 4.6. La reconstruction optimisant le PSNR dans le cas d'un bruit blanc uniformément réparti dans chacune des bandes, est décrite dans [DV03], mais est de complexité légèrement supérieure, pénalisant ainsi l'estimation d'orientation présentée dans la section 4.2.2. Il apparaît que la robustesse des points, des échelles, des orientations, et des descripteurs est peu sensible au bruit de quantification.

5.1.2 Reconstruction par projections sur ensembles convexes

Lors d'une transformée linéaire redondante, l'espace transformé est de plus grande dimension que l'espace image. Considérant la forme vectorisée $\mathbf{x}$ d'une image $I \in \Omega_{nm}$ de $N = n \times m$ pixels, et $\mathbf{y}$ la forme vectorisée des coefficients transformés, l'opérateur d'analyse par une telle transformée peut s'écrire sous forme matricielle A telle que $\mathbf{y} = A\mathbf{x}$. L'opérateur de reconstruction R définie par $R \circ A = Id$ n'est pas unique, et est de la forme :

$$R = R^{\dagger} + U(Id - AR^{\dagger}) \tag{5.1}$$

où $R^{\dagger} = (A^T A)^{-1}$ est le pseudo-inverse de A, et U est une matrice carré quelconque de taille $N \times N$. Le pseudo-inverse est l'opérateur de reconstruction minimisant l'erreur quadratique due à la présence d'un bruit blanc uniformément réparti dans chacune des sous-bandes. Dans [DV03] est montrée que, pour la pyramide laplacienne, la reconstruction classique définie dans la section 4.2 n'est pas optimale, et une méthode de reconstruction par le pseudo-inverse est proposée. Soit $W = A(\Omega_{nm})$ le sous-espace des transformées de tous les signaux $\mathbf{x}$ possibles. Les projections P^W et $P^{\perp}$ respectivement

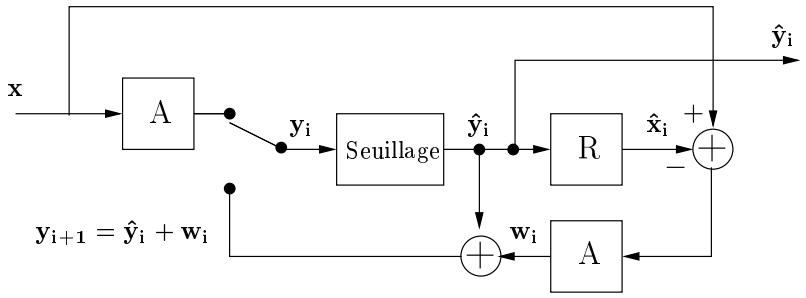

FIG. 5.2 – Schéma de projections itératives sur ensembles convexes.

sur W et sur le supplémentaire orthogonal de W s'écrivent :

$$\begin{aligned} P^W &= AR^\dagger \\ P^\perp &= Id - AR^\dagger \end{aligned}$$

Pour comprendre comment il est possible d'augmenter la parcimonie des représentations redondantes, il est intéressant de décomposer le signal transformé en $\mathbf{y} = P^W\mathbf{y} + P^\perp\mathbf{y}$. La reconstruction par l'équation 5.1 conduit à :

$$\mathbf{x} = R\mathbf{y} = R(P^W\mathbf{y} + P^\perp\mathbf{y}) = RP^W\mathbf{y}$$

Toute modification de $\mathbf{y}$ dans le supplémentaire orthogonal de W ne change pas le signal reconstruit $\mathbf{x}$. Une application directe en compression est de choisir parmi les signaux transformés possibles le plus intéressant, qui peut être le plus creux ou celui d'entropie marginale minimale. Une façon d'effectuer ce choix est de projeter alternativement sur W et sur $W_\mathbf{m}$, l'ensemble des pyramides dont les coefficients sont nuls pour les indices précisés par le masque $\mathbf{m} \in \{0,1\}^M$. Ces deux ensembles sont convexes, assurant la convergence des itérations vers un couple $(\mathbf{y_1}, \mathbf{y_2}) \in W \times W_\mathbf{m}$ dont la distance $||\mathbf{y_1} - \mathbf{y_2}||$ est un minimum local [You78]. Sur la figure 5.2, la pyramide $\mathbf{x}$ ou $\mathbf{y_{i+1}}$ (dont l'inverse est l'image originale) est projetée sur $W_\mathbf{m}$ en $\hat{\mathbf{y}}_\mathbf{i}$. L'opération de sélection des coefficients non nuls par un masque fixe est remplacée par une opération de seuillage. La convexité de $W_\mathbf{m}$ est perdue, mais en pratique la convergence est toujours observée et accélérée. La pyramide $\hat{\mathbf{y}}_\mathbf{i}$ est ensuite projetée en $\mathbf{y_{i+1}}$ sur l'ensemble des pyramides ayant pour inverse l'image originale.

5.1.3 Impact des projections sur la description

Entre deux projections, les coefficients de valeur absolue inférieure à un seuil sont mis à zéro, et l'énergie est redistribuée sur tous les coefficients dans toutes les bandes. Les caractéristiques extraites sont donc susceptibles de modifications importantes entre deux projections. La faible redondance de la transformée limite la redistribution possible d'énergie, et donc la dégradation de la robustesse des caractéristiques extraites au fil des projections. La figure 5.3 montre que, pour une transformation sévère, composée

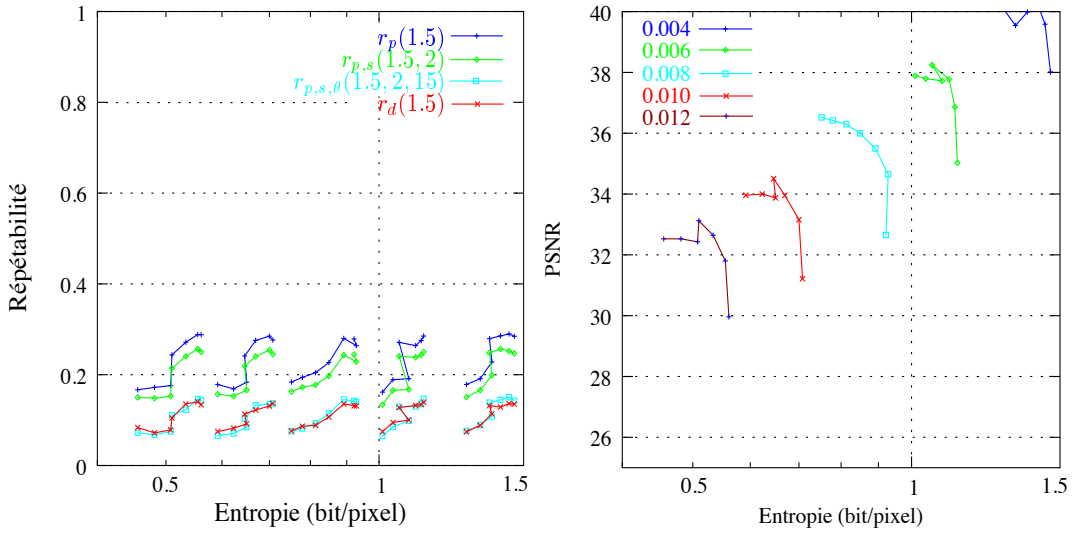

FIG. 5.3 – Répétabilités et PSNR en fonction de l'entropie. Chaque courbe (à droite) ou groupe de courbes (à gauche) est tracée pour un nombre de projections suivant une suite géométrique de 1 à 64, à seuil fixé. De gauche à droite, le seuil définissant les coefficients nuls à chaque projection est donné par la légende de la figure de droite.

d'une dilatation de facteur 1.6, d'une rotation d'angle 30 degrés, d'un crop de facteur 30%, et d'une compression JPEG de facteur 10, il est possible d'obtenir une répétabilité par descripteurs supérieure à 0.15, ce qui est suffisant pour une application en détection de copies. Il est intéressant de constater que les répétabilités décroissent alors que le PSNR croît avec le nombre de projections. Ceci met en évidence une difficulté dans l'élaboration de schémas de compression et de description simultanées. Se pose en effet le problème de l'existence de la bonne mesure de distorsion à minimiser. Les résultats montrent que l'erreur quadratique n'est pas pertinente pour évaluer la dégradation de la description.

5.2 Codage des représentations orientables

Parmi les représentations redondantes étudiées dans le quatrième chapitre, les représentations orientables sont apparues comme les mieux adaptées à la transposition du descripteur SIFT. Elles sont néanmoins fortement redondantes. Comme l'a présenté la section 4.5, la décomposition orientable conduit à la bande correspondant au reste haute fréquence de la relation 4.23 (de taille égale à la taille de l'image), N bandes directionnelles par échelle, et d'une bande correspondant au reste basse fréquence. La redondance de la représentation est donc égale à $1 + \frac{1}{4^{J-1}} + N \sum_{j=0}^{J-1} \frac{1}{4^j} = 1 + \frac{4}{3}NJ$, où J est le nombre d'échelles analysées.

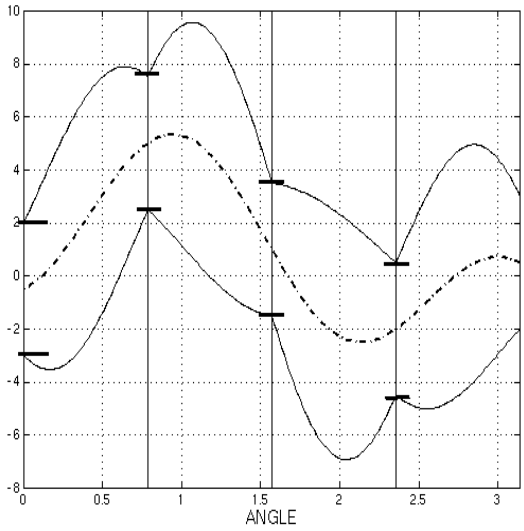

FIG. 5.4 – Région d'incertitude à partir de quatre coefficients quantifiés (extrait de [BLO99]).

5.2.1 Reconstruction contrainte par les fonctions d'interpolation

Une stratégie originale de codage des pyramides orientables est proposée dans [BLO99]. Une représentation orientable définie sur N angles de base est interpolée sur $M > N$ angles. En utilisant correctement les fonctions d'interpolation, un gain de codage est observé en quantifiant non pas la représentation initiale mais la représentation interpolée sur M angles, pourtant de plus grande redondance. Cela est possible par le fait que la quantification de la représentation interpolée peut être rendue beaucoup plus grossière que celle de la représentation initiale. Pour l'expliquer, considérons une représentation orientable définie sur $N = 4$ angles de base $\{\phi_n = \frac{(n-1)\pi}{4}\}_{1\leq n\leq 4}$. À une échelle et une position fixées, les coefficients $c(\phi_n)$ définis sur ces quatre angles permettent de calculer la réponse en un angle ϕ quelconque par le biais des fonctions d'interpolation $\{b_n\}_{1\leq n\leq 4}$ définies par la relation 4.26 (page 117) :

$$c(\phi) = \sum_{n=1}^{4} c(\phi_n) b_n(\phi) \ . \tag{5.2}$$

Si l'on choisit de quantifier la représentation interpolée sur $M > 4$ bandes orientées, le choix d'un seul ensemble de 4 bandes ($\{k_n \in [\![1, M]\!]\}_{1\leq n\leq 4}$) contraint les coefficients des $M - 4$ autres bandes à appartenir à une région d'incertitude définie par :

$$R_M(\phi) = \sum_{n=1}^{4} M(\phi_{k_n}) b_{k_n}(\phi), \quad M(\phi_{k_n}) = \begin{cases} Q_{k_n} \text{ si } b_{k_n}(\phi) > 0 \\ q_{k_n} \text{ si } b_{k_n}(\phi) < 0 \end{cases}$$

$$R_m(\phi) = \sum_{n=1}^{4} m(\phi_{k_n}) b_{k_n}(\phi) \quad m(\phi_{k_n}) = \begin{cases} q_{k_n} \text{ si } b_{k_n}(\phi) > 0 \\ Q_{k_n} \text{ si } b_{k_n}(\phi) < 0 \end{cases}$$

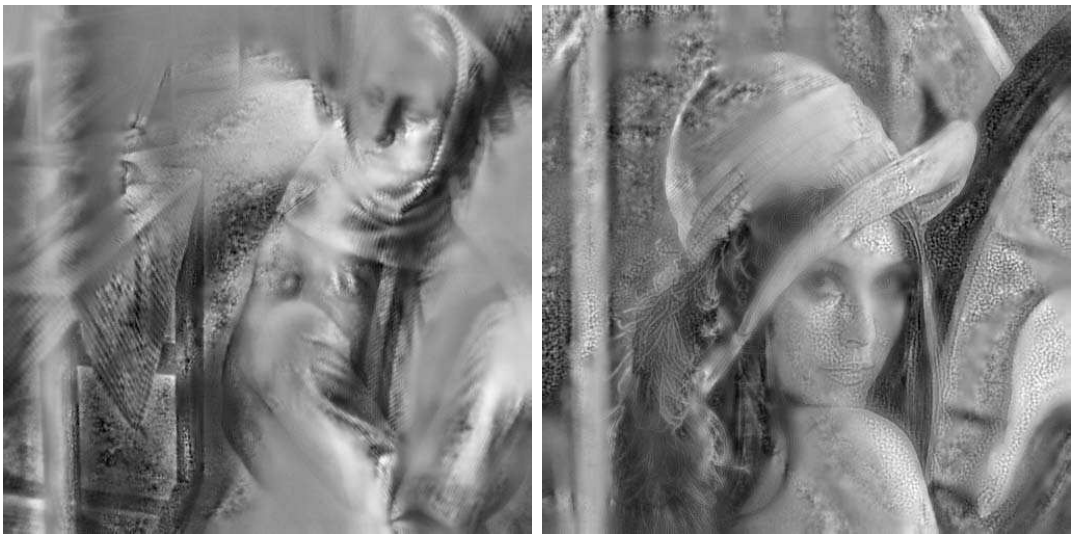

FIG. 5.5 – Importance relative de l'énergie et de l'orientation pour la reconstruction. Reconstruction en combinant l'orientation de l'image Barbara et l'énergie de l'image Lena (à gauche), et en combinant l'orientation de Lena et l'énergie de Barbara (à droite).

où $[q_{k_n}, Q_{k_n}]$ est l'intervalle de quantification du coefficient $c(\phi_{k_n})$. Une région d'incertitude calculée à partir de coefficients quantifiés uniformément est illustrée par la figure 5.4. Le calcul des régions d'incertitude est effectué pour tous les ensembles possibles de 4 bandes. Les coefficients utilisés pour reconstruire l'image sont choisis égaux au centre de l'intersection de toutes les régions d'incertitude (le centre d'un intervalle d'incertitude est représenté en pointillé sur la figure 5.4). Les expériences conduites dans [BLO99] montrent que cette technique permet un gain de codage à bas débit par rapport à la simple quantification des 4 angles de base. Cette technique n'est toutefois pas bien adaptée au problème conjoint de compression et de description, puisque la description requiert l'estimation de l'orientation dominante à partir des coefficients de la pyramide orientable. La quantification grossière ne permet pas une estimation robuste de cette orientation dominante (sauf si l'on calcule l'intersection de toutes les régions d'incertitude, mais cela engendre une description de complexité importante). Dans les techniques présentées ci-dessous, l'orientation dominante est l'information codée, ce qui permet d'effectuer de nombreux traitements visuels directement dans le domaine compressé.

5.2.2 Reconstruction par l'orientation

Dans une représentation orientable à deux bandes orientées (d'angles de base $\phi_1 = 0$ et $\phi_2 = \frac{\pi}{2}$), l'énergie et l'orientation sont définies par :

$$E_s[x,y] = \sqrt{c_{\phi_1}[x,y]^2 + c_{\phi_2}[x,y]^2}, \tag{5.3}$$

$$\Theta_s[x,y] = \text{arctg}(c_{\phi_1}[x,y]/c_{\phi_2}[x,y]). \tag{5.4}$$

Lorsque les coefficients c_{ϕ_1} et c_{ϕ_2} sont les dérivées partielles en x et en y, cette opération revient à exprimer le gradient en coordonnées polaires. Une observation importante est illustrée par la figure 5.5. Les deux images sont reconstruites en inversant l'énergie de l'image Lena avec celle de l'image Barbara. Il apparaît que l'image dont l'orientation est préservée est clairement identifiable sur l'image reconstruite. L'information portée par l'orientation semble beaucoup plus importante pour la reconstruction que celle portée par l'énergie. Cela rejoint un résultat classique selon lequel la phase est plus informative que l'énergie dans le domaine de Fourier [OL81]. Une technique de codage reposant sur l'orientation a récemment été proposée dans [HS05]. La reconstruction s'effectue par projections itératives sur ensembles convexes. Notons Ω_N l'ensemble des images discrètes nulles sur $\mathbb{N}^2 \setminus [\![1, N]\!]^2$, W l'ensemble des représentations orientables, $A : \Omega_N \to W$ la transformation linéaire associant à une image sa pyramide orientable, et $W_N = A(\Omega_N)$. W_N est un sous-espace linéaire, donc convexe, de W. Comme la représentation orientable est un frame ajusté, $A \circ A^\dagger$ est la projection orthogonale de W sur W_N, où $A^\dagger$ est le pseudo-inverse de A. Soit I l'image à compresser, Θ l'union de ses bandes d'orientation Θ_s définies par la relation 5.4, W_Θ l'ensemble des pyramides ayant Θ comme bandes d'orientation, et les mêmes restes basse-fréquence et haute-fréquence que I (telles que définies en introduction de la section). Il est facile de vérifier que W_Θ est convexe. Il est donc possible, par projections alternatives sur W_Θ et W_N, de reconstruire l'image I uniquement à partir de ses restes basse-fréquence et haute-fréquence, et de ses bandes d'orientation. Cette technique montre la possibilité de reconstruire uniquement à partir de l'information d'orientation ; elle n'est en revanche pas applicable en l'état dans un schéma de compression à cause de la lenteur de convergence de la reconstruction par projections (il faut en effet plus de 25 projections pour obtenir un PSNR supérieur à 35).

5.2.3 Reconstruction par l'énergie et l'orientation

La technique précédente reste trop coûteuse en débit et en temps de calcul. Elle nécessite de coder les restes haute-fréquence et basse-fréquence, ainsi que les bandes d'orientation en tout point. Le nombre de coefficients à coder est donc $N(1 + \frac{1}{4^J} + \sum_{j=0}^{J-1} \frac{1}{4^j}) \simeq \frac{7}{3}N$, où N est le nombre de pixels, et J le nombre d'échelles analysées. Outre la redondance élevée, un autre problème est que la distribution des orientations n'est pas creuse. Enfin, la convergence par projection sur les ensembles convexes précédemment définis est très lente.

La solution proposée est composée des étapes suivantes lors de la phase de compression.

1. L'image est transformée en sa représentation orientable sur 4 octaves et sur 2 bandes directionnelles ;
2. La représentation orientable est alternativement projetée sur l'ensemble des représentations à coefficients supérieurs à un seuil en valeur absolue, et sur l'ensemble des représentations ayant pour inverse l'image originale. Les projections

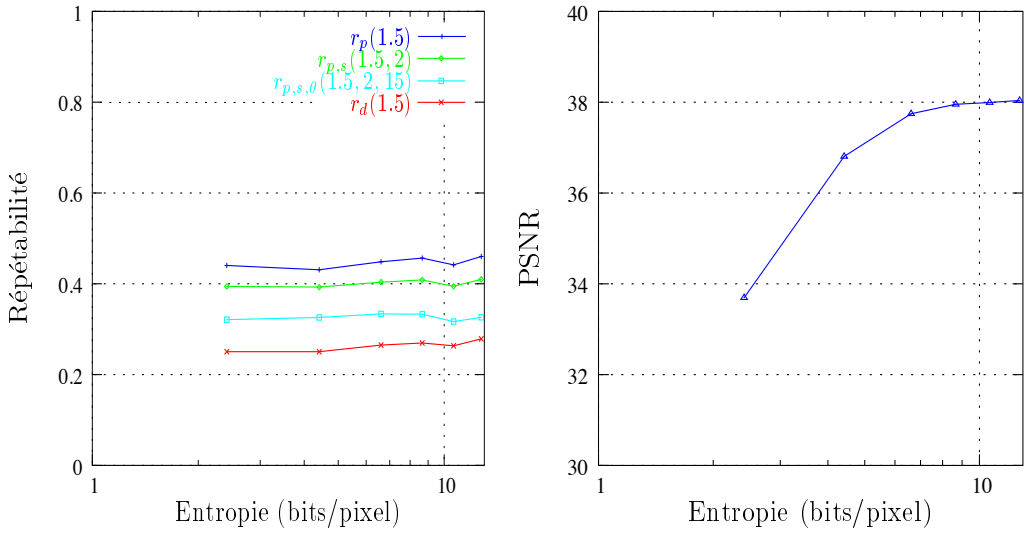

FIG. 5.6 – Impact de la quantification de la bande d'énergie sur la description pour les représentations orientables.

sur ensembles convexes sont utilisées à l'analyse et non pas comme dans la section précédente à la synthèse.

3. À chaque échelle, les deux bandes directionnelles sont transformées en une bande d'énergie et une d'orientation telles que définies par les relations 5.3 et 5.4.
4. La bande d'énergie est quantifié uniformément. Seuls les coefficients de la bande d'orientation ayant une énergie non nulle sont quantifiés. Ils sont également quantifiés uniformément. L'impact des pas de quantification est étudié dans la section 5.2.4.
5. La bande d'énergie et la bande d'orientation sont codées par un codeur entropique de complexité linéaire.

Lors de la phase de décompression, après inversion du codage entropique de la bande d'énergie, les deux bandes directionnelles sont calculées à partir de la bande d'énergie et de la bande d'orientation, puis l'image est reconstruite à partir des bandes directionnelles et des restes basse et haute fréquence.

5.2.4 Impact du bruit de quantification sur la description

Quantification de la bande d'énergie. La bande d'énergie est quantifiée uniformément par pas croissant, allant de 3 à 8 bits. L'orientation est quantifiée par pas de 20 degrés. Rappelons que l'orientation n'est codée qu'aux coefficients d'énergie non nulle. Pour évaluer l'impact de la quantification sur la qualité de la description, il n'y a pas de projections sur ensembles convexes. La figure 5.6 montre que la robustesse des caractéristiques extraites est peu dégradée par la quantification de la bande d'énergie. Cette robustesse n'est pas affectée pour des quantifications allant jusqu'à 3 bits par coefficient.

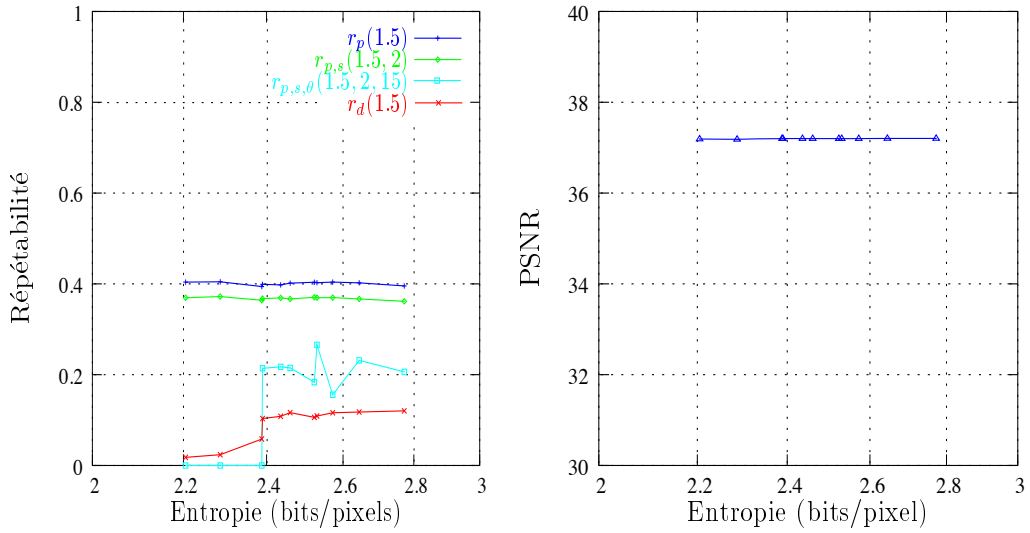

FIG. 5.7 – Impact de la quantification de la bande d'orientation sur la description pour les représentations orientables.

Bien entendu, la redondance de la représentation étant élevée, les entropies ramenées par pixel de l'image originale sont extrêmement élevées.

Quantification de la bande d'orientation. La figure 5.7 montre l'impact de la quantification de la bande d'orientation sur la description (à gauche) et sur la qualité de reconstruction (à droite). La représentation est creusée par projections sur ensembles convexes avant de quantifier, si bien que le nombre de coefficients d'énergie non nulle est faible. Comme seuls ces coefficients sont codés dans la bande d'orientation, l'impact de la quantification en orientation est nul sur la qualité de reconstruction. Sur les deux figures, l'entropie est comprise entre 2.2 et 2.8 bits/pixel, obtenue en quantifiant l'orientation par pas de 5 à 90 degrés. Il est intéressant de constater que la répétabilité $r_{p,s,\theta}(1.5, 2, 15)$ est assez peu sensible aux quantifications par pas 36 degrés. Deux points d'intérêt contribuent à cette répétabilité si l'erreur en orientation est inférieure à 15 degrés. Ceci est possible à cause de la méthode d'estimation qui prend en compte les orientations locales. Il s'agit, en effet, de la méthode reposant sur le maximum de l'histogramme des orientations locales, telle que présentée dans la section 2.3.

5.2.5 Impact du nombre de projections sur la description

Pour obtenir des entropies beaucoup plus faibles, il est possible de creuser la représentation par projections sur ensembles convexes. Le premier ensemble est celui des représentations orientables ayant pour inverse l'image originale. Par linéarité de la transformée, cet ensemble est convexe. Le deuxième ensemble convexe est celui des représentations dont les coefficients sont seuillés en valeur absolue. En fin d'itérations des projections alternatives, l'énergie et l'orientation sont codées d'après la méthode

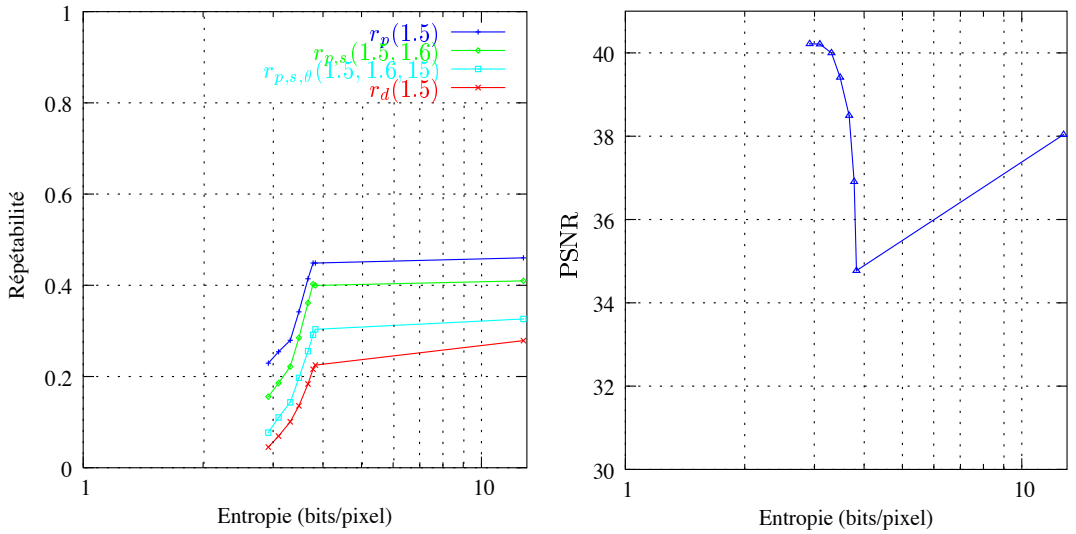

FIG. 5.8 – Impact du nombre de projections sur la description pour les représentations orientables. Ce nombre vaut zéro pour le point le plus à droite, puis suit une suite géométrique de 1 à 64.

proposée dans la section 5.2.4. Le but étant de mesurer la dégradation causée par les projections, l'énergie est codée sur 8 bits par coefficient et l'orientation par pas de 20 degrés. La figure 5.8 montre que la robustesse des caractéristiques est fortement altérée au fil des projections. Ceci provient de la forte redondance de la représentation faisant qu'il existe une large redistribution possible de l'énergie entre deux projections. En particulier, les maxima locaux de l'énergie se déplacent sensiblement, réduisant considérablement la robustesse des points d'intérêt. Il est intéressant de constater que l'impact de la première projection sur la compression et la description est tout à fait différent de l'impact des autres projections. La première projection réduit considérablement l'entropie de 12 à 3.8 bits/pixel à PSNR approximativement constant, et ne modifie presque pas la robustesse des caractéristiques extraites. Les projections suivantes ont un effet très différent : l'entropie ne diminue presque plus et le PSNR croît considérablement, et la robustesse des caractéristiques extraites s'effondre.

5.2.6 Évaluation du schéma complet

Le schéma complet de compression est composé d'une transformation orientable sur deux bandes directionnelles, de projections sur ensembles convexes pour creuser la représentation, du calcul de l'énergie et de l'orientation par les relations 5.3 et 5.4, et de la quantification de l'énergie et de l'orientation aux coefficients d'énergie non nulle.

Les figures 5.10, 5.11, 5.12 aident à fixer trois paramètres : le seuil utilisé dans POCS pour creuser la représentation ; le nombre de projections ; le pas de la quantification de la bande d'énergie. Le seuil de POCS est le même sur les quatre graphiques d'une même page et vaut 0.02, 0.04, et 0.06. Sur une page, la bande d'énergie est quantifiée sur

	Entropie (bits/pixel)	PSNR	Nombre de votes de l'image originale	Deuxième plus grand nombre de votes
Pyramide laplacienne	1.0	35.4	224	41
Représentation orientable	2.4	35.2	33	10

FIG. 5.9 – Résultats de votes pour la pyramide laplacienne et les représentations orientables.

3, 4, 5, et 6 bits de haut en bas et de gauche à droite. Sur un graphique sont tracés les répétabilités et le PSNR obtenus en faisant varier géométriquement le nombre de projections de 1 à 64.

Les observations principales sont les suivantes :

- la quantification a peu d'influence sur les répétabilités ;
- lorsque le seuil de POCS est petit (égal à 0.02), et lorsque la quantification est grossière (sur 3 ou 4 bits), la redistribution d'énergie induite à chaque projection est suffisamment faible pour ne pas perturber les répétabilités (du moins pour un nombre de projections inférieur à 8) ;
- lorsque le seuil de POCS devient grand (supérieur à 0.04), la redistribution d'énergie ayant lieu durant les projections, devient importante et diminue considérablement les répétabilités. Pour un seuil égal à 0.08, il n'est pas possible d'effectuer plus de deux projections.

Le compromis pour le problème conjoint de compression et de description est donc bien moins bon que pour les pyramides laplaciennes. Pour une répétabilité r_d par descripteur de l'ordre de 0.2, et un PSNR de l'ordre de 35, il faut 2.8 bits/pixel dans les représentations orientables (16 projections et quantification sur 4 bits de l'énergie avec un seuil de 0.04, ou 2 projections et quantification sur 5 bits avec un seuil de 0.08) contre 0.6 bits/pixel (8 projections, seuil de POCS=0.01) dans les pyramides laplaciennes.

5.3 Détection de copies

Cette section présente les résultats de détection de copies dans le domaine compressé. Rappelons que le domaine compressé fait référence au domaine quantifié, ce qui implique l'inversion du codeur arithmétique. Le schéma de détection est présenté en introduction par la figure 2 (page 16). Ce schéma consiste, pour chaque descripteur local d'une image requête (l'image pour laquelle il faut décider s'il s'agit d'une image de la base), à chercher le descripteur le plus proche parmi tous les descripteurs de la base. Ce nombre est ici important, égal à 3.10^4 images $\times 10^3$ descripteurs/image $= 3.10^7$ descripteurs. L'image possédant le descripteur le plus proche vote une fois. Les images sont finalement classées par nombre décroissant de votes. Une suspicion de copie apparaît alors si le nombre de votes de la première image est significativement au-dessus du niveau de bruit.

La figure 5.3 donne les résultats de votes pour les pyramides laplaciennes et les représentations orientables. Les résultats sont obtenus par moyenne sur vingt requêtes aléatoirement extraites parmi les 30 000 images de la base. La copie est créée par une dilatation de facteur 1.6, une rotation de 30 degrés, une compression JPEG de facteur 10, et d'un crop de facteur 30%. La recherche de l'attaque optimale, c'est-à-dire limitant au plus le nombre de votes de l'image originale pour une distorsion donnée, n'a pas été traitée. Les résultats de détection de copie sont encourageants. Dans la quasi totalité des cas, le nombre de votes de l'image originale était suffisamment important pour suspecter l'image requête comme copie. Dans seulement deux cas, l'image originale n'a pas été l'image votant le plus. Ces cas sont apparus pour la représentation orientable. Dans les deux cas, l'image ayant voté le plus était une image de la même scène que l'image originale.

5.4 Conclusion

Ce chapitre a proposé deux schémas de compression permettant l'extraction de descripteurs dans le domaine quantifié. Ces deux schémas reposent respectivement sur les représentations laplaciennes et sur les représentations orientables. Les deux types de représentation étant redondantes, il est possible d'utiliser les projections sur ensembles convexes pour trouver une représentation plus creuse. Les tests ont montré que les projections sont à utiliser avec précaution, surtout si la redondance de la représentation est élevée. En effet, dans ce cas, la redistribution d'énergie peut être importante et modifier de façon significative les caractéristiques extraites. D'autres tests ont montré la faible sensibilité de ces caractéristiques à la quantification. Une technique de codage originale à également été proposée pour les représentations orientables. Pour un PSNR de 35, l'entropie reste néanmoins beaucoup plus élevée pour ce type de représentations que pour les pyramides laplaciennes (2.4 bits/pixel à comparer à 1.0 bit/pixel pour les tests réalisés en détection de copies). Toutefois, la complexité de description dans les représentations orientables est beaucoup plus faible. L'information d'énergie et d'orientation nécessaire pour la description est en effet directement accessible dans le domaine compressé.

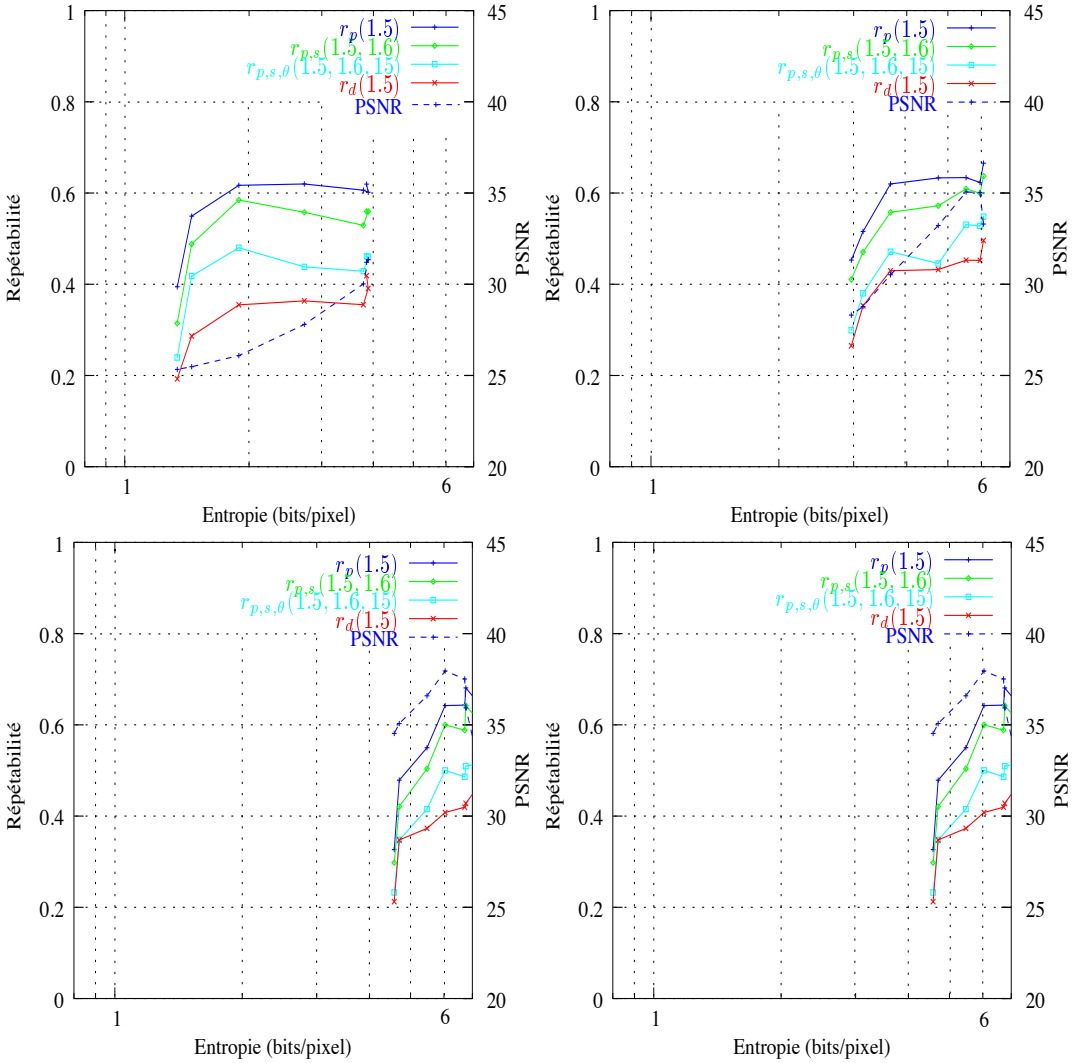

FIG. 5.10 – POCS à l'analyse (seuil=0.02), entre 1 et 64 projections, suivie d'une quantification uniforme sur 3 bits (en haut à gauche), 4 bits (en haut à droite), 5 bits (en bas à gauche), et 6 bits (en bas à droite).

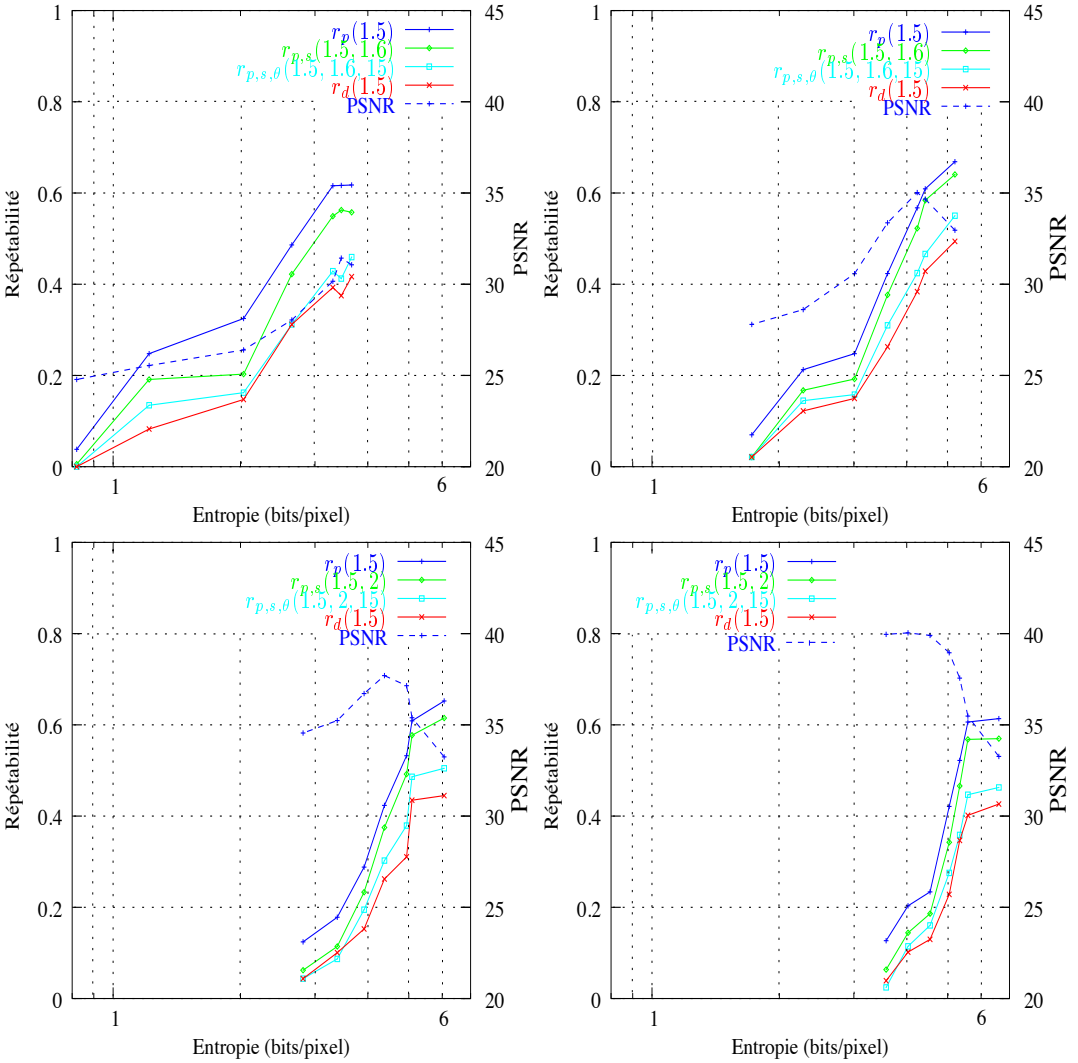

FIG. 5.11 – POCS à l'analyse (seuil=0.04), entre 1 et 64 projections, suivie d'une quantification uniforme sur 3 bits (en haut à gauche), 4 bits (en haut à droite), 5 bits (en bas à gauche), et 6 bits (en bas à droite).

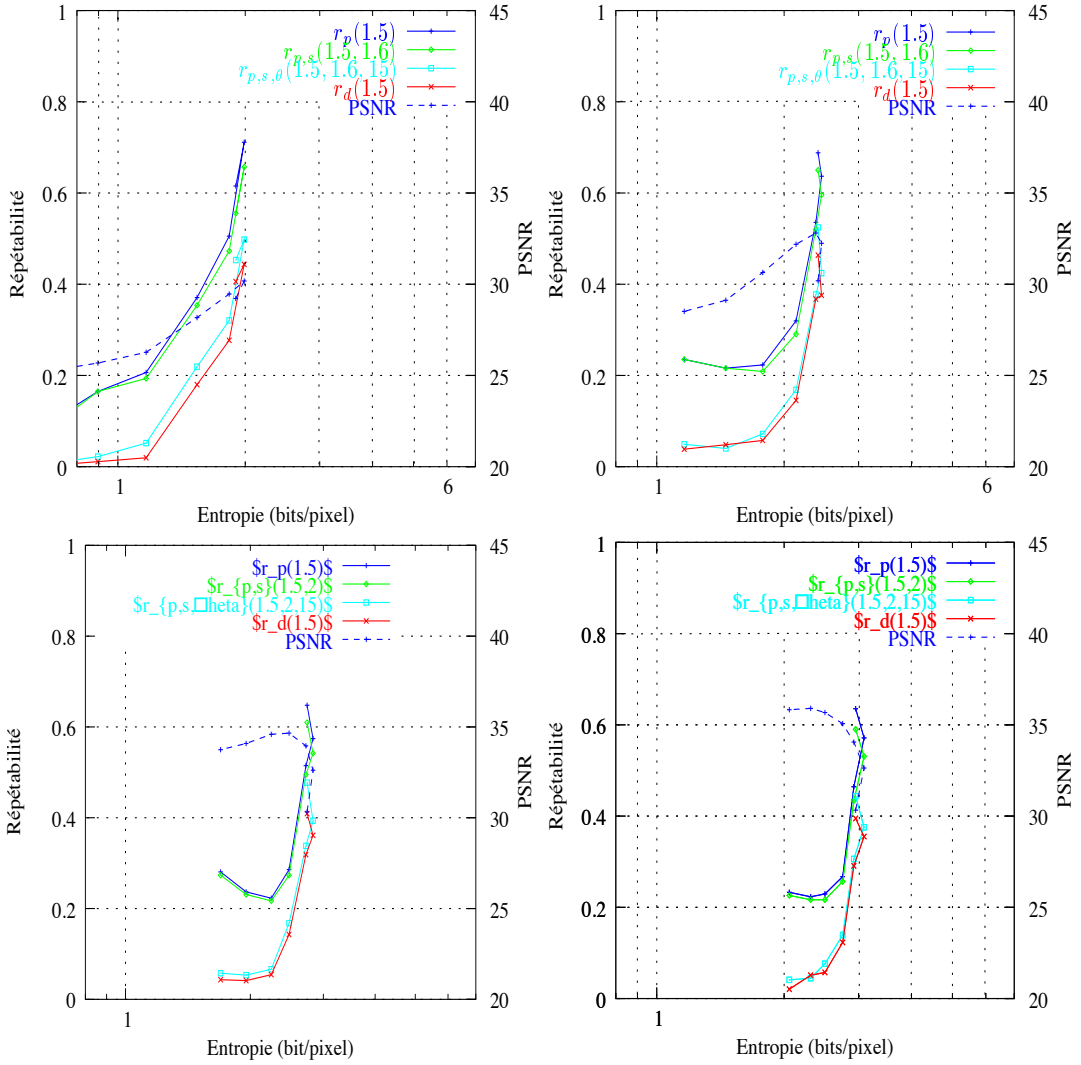

FIG. 5.12 – POCS à l'analyse (seuil=0.08), entre 1 et 64 projections, suivie d'une quantification uniforme sur 3 bits (en haut à gauche), 4 bits (en haut à droite), 5 bits (en bas à gauche), et 6 bits (en bas à droite)

Conclusion

Le problème central dans les chaînes de communication d'images était jusqu'à une époque récente la transmission sur des canaux contraints en débit. L'accroissement des bandes passantes et des espaces de stockage relâche sensiblement cette contrainte, mais en fait apparaître une autre : celle de transmettre à l'utilisateur la *bonne* image parmi un grand ensemble d'images stockées. Cette nouvelle contrainte nécessite des traitements de vision artificielle pour comprendre le besoin de l'utilisateur (homme ou machine). Cette thèse constitue un travail préliminaire pour l'émergence d'un standard de compression permettant la réalisation de ces traitements visuels en temps réel dans le domaine compressé. Les principales conclusions et perspectives de cette investigation sont exposées dans les deux prochaines sections

5.5 Conclusions

Nécessité d'évaluer la complexité des schémas actuels de description. Les critères d'évaluation sont typiquement des taux de bonne et mauvaise détection en classification d'objets, de groupes d'objets, ou de scènes, et ne prennent pas en compte la complexité. Parmi les deux méthodes de référence, respectivement issues du modèle probabiliste de constellation présenté dans [FPZ03] et du modèle déterministe de Lowe [Low99], la complexité de la première exclut toute application temps réel, et celle de la seconde peut être significativement diminuée avec une faible dégradation de performance. Le second chapitre a ainsi montré que, certes une représentation multi-échelles est nécessaire pour la description, mais que la discrétisation en échelle peut être beaucoup plus faible que celle classiquement utilisée [Low04]. Dans les techniques de description, la prise en compte de la contrainte de complexité est nécessaire pour l'émergence d'applications temps réel dans les chaînes de communication d'images.

Nécessité de décompression partielle avant de décrire. L'élaboration de schémas de description peu complexe est d'autant plus importante qu'il n'existe aucun codage préservant l'information visuelle, rendant nécessaire l'étape de décodage pour l'exécution de traitements visuels. À défaut d'obtenir de tels codeurs, il est possible d'utiliser un codeur arithmétique de complexité linéaire, bien inférieure à celle des codeurs de référence comme JPEG 2000. Dans le cas courant des représentations multi-échelles,

il peut être nécessaire de reconstruire les bandes d'approximation avant d'extraire l'information visuelle pertinente. La complexité de la description doit alors être réduite au maximum.

Nécessité de l'évaluation simultanée du débit, de la qualité des images reconstruites, et de la qualité de la description. Les schémas de compression sont guidés par l'optimisation de la fonction débit-distorsion. Le choix du PSNR comme mesure de distorsion est valable pour mesurer la qualité des images reconstruites, pas pour mesurer la capacité à décrire dans le domaine transformé (qui peut s'apparenter à la covariance de l'image transformée). La définition d'une mesure de distorsion prenant en compte ces deux aspects permettrait de dériver théoriquement l'allocation optimale du débit au sens d'un certain compromis entre qualité visuelle et qualité de description. À défaut d'une telle définition, il est possible de mesurer le compromis pouvant être atteint expérimentalement. La détection de copies constitue un possible cadre d'évaluation. Par exemple, à débit fixé, les performances dépendent du PSNR et des taux de correctes et fausses détections, que l'on peut mesurer en faisant varier la taille de la base et la sévérité de la transformation.

Caractérisation de la représentation d'image adaptée au problème conjoint de compression et description. Les représentations d'images utilisées en compression satisfont rarement la contrainte de covariance aux transformations géométriques. Cette contrainte est pourtant nécessaire pour la réalisation de la grande majorité des traitements visuels. Seule la description globale peut se passer de cette contrainte, grâce à son effet moyennant sur toute l'image. Le premier chapitre a montré que les représentations (linéaires) covariantes sont multi-échelles isotropes, ou multi-échelles et multi-orientations. Le premier type de représentation implique une plus grande complexité de description, puisqu'il requiert la reconstruction des bandes d'approximation. Le quatrième chapitre a mis en évidence les points exposés ci-dessous.

- La transformée en ondelettes non sous-échantillonnée [Dut89] a montré qu'un échantillonnage dyadique en échelle est suffisant pour maintenir une bonne qualité de description.
- La pyramide laplacienne [BA83] a montré qu'un sous-échantillonnage spatial inter-échelle est possible, il suffit par exemple de ne pas sous-échantillonner la bande de détails.
- La représentation en contourlets [Do01] est fortement variante aux rotations, si bien que l'analyse directionnelle n'est pas exploitable pour des traitements visuels. Ce résultat est prévisible puisque le banc de filtres directionnels est a échantillonnage critique.
- La représentation en ondelettes complexes [Kin98] est peu variante aux rotations, mais fortement aux translations à cause d'un trop fort échantillonnage spatial.
- La représentation orientable est peu variante à toutes les transformations géométriques. Elle est consiste en la mise en cascade d'une pyramide laplacienne et d'un

banc de filtres orientables non sous-échantillonne sur la bande de détails. Cette représentation présente donc de nombreuses possibilités en traitement visuel, mais est fortement pénalisée en compression par sa redondance importante.

Effets de la compression sur la description. Le dernier chapitre a montré que la qualité de description résiste bien à la quantification des coefficients transformés. La technique POCS classiquement utilisée pour augmenter la parcimonie des représentations redondantes est en revanche à manier avec prudence. Pour les représentations faiblement redondantes comme les pyramides laplaciennes, cette technique a peu d'impact sur la qualité de description. Il en va tout autrement des représentations fortement redondante comme les représentations orientables, où la description est considérablement dégradée au-delà de quelques projections. Un schéma de description de faible complexité (et ne requérant pas la reconstruction des bandes d'approximation) a été proposé à partir de représentations orientables. La qualité de description est apparue suffisante pour permettre la détection de copies à partir d'une base de 30 000 images.

5.6 Perspectives

Le travail d'investigation mené dans cette thèse conduit à plusieurs perspectives.

- Il doit exister une adaptation entre le descripteur et la perte d'information due aux techniques de compression. Pour la technique POCS par exemple, le problème est de savoir quelle est la description la plus résistante, ou de choisir les ensembles convexes minimisant la dégradation de la description.
- Si la description requiert la reconstruction des bandes d'approximation, il est possible de commencer le traitement visuel avec les descripteurs calculés à partir des bandes de résolution grossière. Dans le cas de la détection de copies, le système de votes peut être considérablement accéléré par cette description hiérarchique. Les premiers descripteurs calculés peuvent servir à sélectionner les images de la base d'où est le plus vraisemblablement issue la copie.
- Les schémas de description locale peuvent s'étendre à d'autres applications que la détection de copie. Une question est de savoir quelle performance de reconnaissance d'objets ou de scènes il est possible d'atteindre dans le domaine compressé.
- Les représentations d'images covariantes aux transformations géométriques sont des candidates naturelles pour l'estimation dense de mouvement entre images d'une vidéo. Dans le standard MPEG, le mouvement s'estime entre blocs, si bien que les erreurs d'appariement entre blocs perturbent considérablement les techniques possibles de description. Un standard de vidéo où l'estimation de mouvement est dense et multi-échelle permettrait des descriptions plus riches et plus fiables.

Bibliographie

[ADJ+02] J.P. Antoine, L. Demanet, L. Jacques, J.F. Hochedez, R.Terrier, et E. Verwichte. Applications of the 2-d wavelet transform to astrophysical images. *Physicalia magazine*, 24:93–116, 2002.

[BA83] P. J. Burt et E. H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 9(4):532–540, 1983.

[BAG03] S.A. Berrani, L. Amsaleg, et P. Gros. Robust content-based image searches for copyright protection. Dans *Proceedings of the ACM International Workshop on Multimedia Databases*, pages 70–77, La Nouvelle Orléans, Louisiane, Novembre 2003.

[Bea78] P.R. Beaudet. Rotational invariant image operators. Dans *Proceedings of the Fourth International Conference on Pattern Recognition*, pages 579–583, Tokyo, Japon, 1978.

[Big04] T. Bigun. Recognition by symmetry derivatives and the generalized structure tensor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):1590–1605, 2004.

[BLO99] B. Beferull-Lozano et A. Ortega. Coding techniques for oversampled steerable transforms. Dans *Proceedings of the Thirty-Third International Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1198–1202, Pacific Grove, CA, USA, Octobre 1999.

[BMP02] S. Belongie, J. Malik, et J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Analysis*, 24(24):509–522, 2002.

[BS97] A.J. Bell et T.J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[CB05] R. Coudray et B. Besserer. Agrégation, sélection et utilisation de l'information de mouvement issue d'un flux MPEG. Dans *Vingtième Colloque GRETSI sur le Traitement du Signal et des Images*, pages 1161–1164, Louvain-la-Neuve, Belgique, Septembre 2005.

[CD99] E. Candès et D. Donoho. *Curves and Surfaces*, chapitre Curvelets: A surprisingly effective nonadaptive representation of objects with edges. Vanderbilt University Press, 1999.

[Chi03] T.T. Chinen. Visual comparison of JPEG 2000 versus conventional JPEG. Dans *Proceedings of the IEEE International Conference on Image Processing*, pages 283–286, Barcelone, Septembre 2003.

[CJ02] G. Carneiro et A.D. Jepson. Phase-based local features. Dans *Proceedings of the Seventh European Conference on Computer Vision*, volume 1, pages 282–296, Mai 2002.

[CLS95] C.H. Chen, J.S. Lee, et Y.N. Sun. Wavelet transformation for gray level corner detection. *Pattern Recognition Letters*, 28(6):853–861, 1995.

[Dau92] I. Daubechies. *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[DF90] R. Deriche et O. Faugeras. 2-D curve matching using high curvature points: application to stereo. Dans *Proceedings of Tenth International Conference on Pattern Recognition*, pages 18–23, Atlantic City, USA, Juin 1990.

[Do01] M.N. Do. *Directional MultiResolution Image Representations*. PhD thesis, Swiss Federal Institute of Technology, Lausanne, Switzerland, 2001.

[Dut89] P. Dutilleux. *Wavelets: time-frequency methods and phase-space*, chapitre An implementation of the "algorithme a trous" to compute the wavelet transform, pages 298–304. Springer-Verlag, 1989.

[DV00] M.N. Do et M. Vetterli. Orthonormal finite ridgelet transform for image compression. Dans *Proceedings of the IEEE International Conference on Image Processing*, pages 367–370, Vancouver, Canada, Septembre 2000.

[DV02] M.N. Do et M. Vetterli. Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance. *IEEE Transactions on Image Processing*, 11(2):146–158, 2002.

[DV03] M.N. Do et M. Vetterli. Framing pyramids. *IEEE Transactions on Signal Processing*, 51(9):2329–2342, 2003.

[dWSD99] G.V. de Wouwer, P. Scheunders, et D. Van Dyck. Statistical texture characterization from discrete wavelet representations. *IEEE Transactions on Image Processing*, 8(4):592–598, April 1999.

[dWSLD99] G.V. de Wouwer, P. Scheunders, S. Livens, et D. Van Dyck. Wavelet correlation signatures for color texture characterization. *Pattern Recognition*, 32:443–451, 1999.

[FA91] W.T. Freeman et E.H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[Far92] M. Farge. Wavelets transforms and their applications to turbulence. *Annual Review on Fluid Mechanics*, 24:395–457, 1992.

[FB04] F. Fraundorfer et H. Bischof. Evaluation of local detectors on non-planar scenes. Dans *Twenty Eighth Workshop of the Austrian Association for Pattern Recognition*, pages 125–132, Hagenberg, 2004.

[Fie87] D.J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Optical Society of America*, 4(12):2379–2394, 1987.

[FPZ03] R. Fergus, P. Perona, et A. Zisserman. Object class recognition by unsupervised scale-invariant learning. Dans *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03)*, volume 2, pages 264–271, Madison, WI, USA, Juin 2003.

[Gho94] F. Ghorbel. A complete invariant description for gray-level images by the harmonic analysis approach. *Pattern Recognition Letters*, 15:1043–1051, 1994.

[GK95] G. Granlund et H. Knutsson. *Signal Processing for Computer Vision*. Kluwer, 1995.

[Gop03] R.A. Gopinath. The phaselet transform - an integral redundancy nearly shift-invariant wavelet transform. *IEEE Transactions on Signal Processing*, 51(7):1792–1805, 2003.

[Gro86] A. Grossmann. Wavelet transform and eedge detection. *Stochastic Processes in Physics and Engineering*, pages 149–157, 1986.

[HKMMT89] M. Holschneider, R. Kronland-Martinet, J. Morlet, et P. Tchamitchian. *Wavelets: time-frequency methods and phase-space*, chapitre A real-time algorithm for signal analysis with the help of the wavelet transform, pages 286–297. Springer-Verlag, 1989.

[HS88] C. Harris et M. Stephens. A combined corner and edge detector. Dans *Proceedings of Fourth Alvey Vision Conference*, pages 147–151, Manchester, Royaume Uni, 1988.

[HS05] D.K. Hammond et E.P. Simoncelli. Nonlinear image representation via local multiscale orientation. Rapport technique, Courant Institute of Mathematical Sciences, New York University, 2005. TR2005-875.

[Hu62] M.K. Hu. Visual pattern recognition by moment invariants. *IEEE Transations on Information Theory*, 8:179–187, 1962.

[HW62] D. Hubel et T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–154, 1962.

[JH99] A. E. Johnson et M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):433–449, 1999.

[JSIN06] J.Malo, E.P. Simoncelli, I.Epifanio, et R. Navarro. Non-linear image representation for efficient perceptual coding. *IEEE Transactions on Image Processing*, 15(1):68–80, 2006.

[KB01] T. Kadir et M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[KC92] A. Kundu et J.L. Chen. Texture classification using QMF bank-based subband decomposition. *Graphical Model and Image Processing*, 54(5):369–384, 1992.

[KD87] J.J. Koenderinck et A.J. Van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55(6):367–375, Mars 1987.

[Kin98] N. Kingsburry. The dual-tree complex wavelet transform: a new efficient tool for image restoration and enhancement. Dans *Proceedings of the European Signal Processing Conference, EUSIPCO 98*, pages 319–322, Septembre 1998.

[Koe84] J.J. Koenderinck. The structure of images. *Biological Cybernetics*, 50:363–370, 1984.

[Kov97] P. Kovesi. Symmetry and asymmetry from local phase. Dans *Proceedings of Tenth Australian Joint Conference on Artificial Intelligence (AI'97)*, pages 185–190, 1997.

[KR82] L. Kitchen et A. Rosenfeld. Gray-level corner detection. *Pattern Recognition Letters*, 1(2):95–102, 1982.

[KR03] N.G. Kingsbury et T. Reeves. Iterative image coding with overcomplete complex wavelet transforms. Dans *Proceedings of the Conference on Visual Communications and Image Processing*, pages 1253–1264, Lugano, Suisse, Juillet 2003.

[KS04] Y. Ke et R. Sukthankar. PCA-SIFT: a more distinctive representation for local image descriptors. Dans *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 506–513, juin 2004.

[Lej05] H. Lejsek. The pvs-index. Thèse de Master, Reykjavik University, 2005.

[Lin94a] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, pages 79–116, 1994.

[Lin94b] T. Lindeberg. *Scale-space theory in computer vision.* Kluwer Academic Publisher, 1994.

[Lin97] T.W. Lin. Compressed quadtree representations for storing similar images. *Image and Vision Computing*, 15:833–843, 1997.

[Lip69] M. Lipschutz. *Differential Geometry*, page 234. McGraw-Hill, 1969.

[Low99] D.G. Lowe. Object recognition from local scale-invariant features. Dans *International Conference on Computer Vision*, pages 1150–1157, 1999.

[Low04] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[LS99] E. Loupias et N. Sebe. Wavelet-based salient points for image retrieval. Rapport Technique RR 99.11, Laboratoire Reconnaissance de Formes et Vision, INSA Lyon, 1999.

[Mal89] S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 674–693, 1989.

[Mal91] S. Mallat. Zero-crossings of a wavelet transform. *IEEE transactions on Information Theory*, 37(4):1019–1034, 1991.

[Mal98] S. Mallat. *A Wavelet Tour of Signal Processing*, page 41. Academic Press, 1998.

[MG01] D. Mumford et B. Gidas. Stochastic models for generic images. *Quaterfly of Applied Mathematics*, LIV(1):85–111, 2001.

[MH92] S. Mallat et W.L. Hwang. Singularity detection and processing with wavelets. *IEEE Transactions on information Theory*, 38(2):617–643, 1992.

[MHS05] E.N. Mortensen, H.D., et L.G. Shapiro. A SIFT descriptor with global context. Dans *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 184–190, San Diego, CA, USA, Juin 2005.

[Mor77] H.P. Moravec. Towards automatic visual obstacle avoidance. Dans *Proceedings of International Joint Conference on Artificial Intelligence*, page 584, Aout 1977.

[MS05] K. Mikolajczyk et C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[MZ93] S. Mallat et Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[Nei64] U. Neisser. Visual search. *Scientific American*, 210(6):94–102, 1964.

[OF97] B. Olshausen et D. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1 ? *Vision Research*, 37:3311–3325, 1997.

[OL81] A.V. Oppenheim et J.S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69:529–541, Mai 1981.

[PAK99] F.A.P. Petitcolas, R.J. Anderson, et M.G. Kuhn. Information hiding – a survey. *IEEE Special Issue on "Identification and protection of multimedia information"*, 87(7):1062–1078, 1999.

[Res95] D. Resifeld. Context free attentional operators: the generalized symmetry transform. *International Journal of Computer Vision*, 14:119–130, 1995.

[SAH92] E.P. Simoncelli, W.T. Freeman E.H. Adelson, et D.J. Heeger. Shiftable multiscale transforms. *IEEE Transactions on Information Theory*, 38(2):587–607, March 1992.

[SB91] M. Swain et D. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[SC96] B. Schiele et J.L. Crowley. Object recognition using multidimensional receptive field histograms. Dans *Proceedings of the Fourth European Conference on Computer Vision (ECCV'96)*, pages 50–54, Cambridge, Royaume Uni, Avril 1996.

[SC00] B. Schiele et J.L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000.

[SCD02] J.L. Starck, E. Candes, et D. Donoho. The curvelet transform for image denoising. *IEEE Transactions on Image Processing*, 11(6):670–684, 2002.

[Sel04] I. W. Selesnick. The double-density dual-tree dwt. *IEEE Transactions on Signal Processing*, 52(5):1304–1314, 2004.

[SFM] J.L. Starck, J. Fadili, et F. Murtagh. The undecimated wavelet decomposition and its reconstruction. submitted.

[Sha93] J.M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41(12):3445–3462, 1993.

[SM96] C. Schmid et R. Mohr. Image retrieval using local characterization. Dans *Proceedings of the International Conference on Communicating by Image and Multimedia*, pages 45–50, May 1996.

[SM97] C. Schmid et R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534, May 1997.

[SMB98] C. Schmid, R. Mohr, et C. Bauckhage. Comparing and evaluating interest points. Dans *International Conference on Computer Vision*, pages 230–235, 1998.

[SR96] B.C. Smith et L.A. Rowe. Compressed domain processing of JPEG-encoded images. *Real-Time Imaging*, 2(1):3–17, 1996.

[VBLVD] V. Velisavljevic, B. Beferull-Lozano, M. Vetterli, et P.L. Dragotti. Directionlets: anisotropic multi-directional representation with separable filtering. *IEEE Transactions on Image Processing.* Accepté en Septembre 2005.

[VJ02] G. Voulgaris et J. Jiang. Quadtree based image indexing in wavelets compressed domain. Dans *Proceedings of the Twentieth Eurographics UK Conference*, pages 89–94, 2002.

[Wit83] A.P. Witkin. Scale space filtering. Dans *International Joint Conference on Artificial Intelligence*, pages 1019–1022, 1983.

[WWP00] M. Weber, M. Welling, et P. Perona. Unsupervised learning of models for recognition. Dans *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 18–32, 2000.

[YF02] S.M. Yamany et A.A. Farag. Surfacing signatures: an orientation independent free-form surface representation scheme for the purpose of objects registration and matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1105–1120, 2002.

[You78] D.C. Youla. Generalized image restoration by the method of alternating orthogonal projections. *IEEE Transactions on Circuits and Systems*, 25(9):694–702, 1978.